DOCUMENT RESUME

ED 269 450 TM 860 284

AUTHOR Allerup, Peter
TITLE Why I Like to Read--Statistical Analysis of Questionnaire Data.
INSTITUTION Danish Inst. for Educational Research, Copenhagen.
REPORT NO DPI-1985-61; ISBN-87-7414-223-3
PUB DATE 85
NOTE 64p.
PUB TYPE Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Data Collection; Elementary Education; Factor Analysis; Foreign Countries; Foreign Students; Goodness of Fit; Latent Trait Theory; Mathematical Models; *Models; Questionnaires; *Reading Attitudes; Scoring; *Statistical Analysis; Statistical Studies; Testing
IDENTIFIERS Denmark; *Rasch Model

ABSTRACT

Approximately 1200 Danish school children aged 8, 11, and 14 participated in a questionnaire study under the caption, "Why I like to read." Data from Denmark formed part of an international study, in which 12 other countries used the same questionnaire to explore various functions of reading and to compare them over different countries. This study concentrated on gathering information by questionnaires, using the "why I like to read" data as examples. The 50 questions are described as dividing points for reducing differentiated information, given in a whole series of questions, down to score information. Such information is calculated based on components resulting from factor analyses indicating a group of questions which belongs to one conceptual unity related to reading. These components are grouped into seven dimensions: enjoyment-stimulation, enjoyment-fiction, future, diversion, my country, professional skill, and improvement of reading. A class of statistical models, especially designed to deal with categorical data, is used for these analytic purposes. The "why I like to read" data is analyzed from the viewpoint of Rasch models and is carried through testing procedures, resulting in homogeneous items. Misfits to the model are considered to be useful conflicts on which to base a new analysis. The appendix contains the questionnaire. (LMO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED269450

Why I like to read - statistical analysis of questionnaire data

Peter Allerup

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Kirsten

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Danish National Institute for Educational Research. 1985

TM 860 284

Why I like to read
- statistical analysis of questionnaire data

Peter Allerup

The Danish National Institute for
Educational Research. 1985

Tryk: DPI off-set
Renskrift: Kirsten Falkenberg and Hanne Ankerdal
DPI nr. 1985.61
ISBN nr. 87 7414 223 3

CONTENTS

1. Introduction. 2
2. Basic notations in Rasch models. 4
3. Responses-data 5
4. The Rasch model 6
5. Test of fit. 7
6. Interpretations of tests 10
7. Some summary statistics 12
8. How to define components. 17
9. What can be assessed by Rasch models? 20
10. The scoring system. 21
11. DIMENSION 1: Enjoyment 26
12. DIMENSION 2: Future 28
13. DIMENSION 3: Diversion 30
14. DIMENSION 4: My country. 32
15. DIMENSION 5: Professional skill. 34
16. Analyses of homogeneity. 36
17. The search for homogeneity. 38
18. Twisting items 42
19. Factor analysis and homogeneity. 45
20. Controlled and uncontrolled analyses. 49
21. Conclusions. 53
22. References 54
23. Appendix 1. 56

1. Introduction.

Approximately 1200 Danish school children, aged 8, 11 and 14, have participated in a questionnaire study under the caption: why I like to read. The translation into Danish from the original questionnaire in English was made by Mogens Jansen and Lotte Rosbaek Juhl.

The collection in Denmark was supervised by Mogens Jansen with valuable assistance from a series of the concultants in Danish (mothertongue teaching) throughout the country.

The provisionally processed data were presented by Peter Allerup at a seminar at the Danish National Institute for Educational Research in the autumn of 1984, in which, among others, several concultants of reading and Danish took part.

Data were collected during school hours and sufficient time was given to make it possible even for the eight-year-old children to complete the questionnaire themselves.

The data from Denmark form part of an international study conducted by Vincent Greaney and Susan Neuman, in which data from twelve other countries are collected by means of exactly the same questionnaire as used in Denmark. Their first analyses of data are presented earlier (Greaney and Neuman, 1984).

While the international study is heading in the direction of making explorative studies of various functions of reading and comparing such functions over different countries (culture, educational tradition...etc.) we will concentrate on certain aspects of the way in which information is gathered by means of questionnaires.

It means in the first place, that we will discuss how those fifty questions are suited for 'measuring purposes', i.e. how they act as dividing points on a ruler if one wants to 'squeeze' the differentiated information given in a whole series of questions down to 'score' information. Such information is calculated based on 'components' (Greaney and Neuman, 1984) resulting from factor analyses or 'dimensions', indicating a group of questions which belongs to one conceptual unity related to reading (to be defined later).

We shall demonstrate how a class of statistical models, especially designed to deal with categorical data can be used for these analytic purposes. The so-called Rasch model will be presented and emphasis given to the philosophy behind and applications of the model rather than to algebra connected with estimation, test...etc. To this end interested readers are referred to the list of references, where you will find original works by G. Rasch (Rasch, 1960, 1961, 1968, 1977), a comphrehensive reading (Fischer, 1974) in which Rasch models are discussed in relation to classical statistical methods and a series of other works with independent usage of Rasch models or works relating Rasch models to modern latent trait theory (e.g. Andersen, 1983, Andersen and Madsen, 1976).

To avoid repeated definitions of basic concepts in Rasch models during the analysis of data, we have devoted a special section to 'basic notations

in Rasch models'. It is meant to serve as a kind of summary but contains also fragments of arguments, a little mathematics and interpretational specialities for these models.

It is not to say that all planned calculations by means of Rasch models are ready until now. In fact, a series of analyses based on what one could call 'Rasch dimensions' remains. This report is intended, primarily, to give an idea of what can be gained by analysing the 'why I like to read'-data from the viewpoint of Rasch models and to go through the testing procedures resulting in homogeneous items. A group of homogeneous items, i.e. Rasch-model-fitted items, then constitutes what we call a 'dimension'! Designed to measure a specific aspect of reading.

It will be clear from the reading that analyses by means of Rasch models are **not** following the otherwise well-known scheme, in which data are handed over to a statistician who in turn fight a computer in order to get data 'analysed' in a statistical package. On the contrary, one could say Rasch model analyses is a **process** of interactions between statistical consequences and psychological evaluations in which much of the hypothesis testings are in fact runned during the process towards 'final' results.

This is why the report is mainly dealing with 'discouraging' results or 'misfits' of the model to data. Misfits are usefull conflicts to base new steps of analysis upon if the instrument, the model creating the misfit provides clear understanding of 'what was wrong'.

2. Basic notations in Rasch models.

Frame of reference refers to the 'experimental design' in which persons (here pupils) give answers to questions (items), and in which a person's response behaviour is fully characterized by one latent parameter σ_v, a theoretical number to be estimated. Interpretation of this number depends on the content of questions. If, e.g. in a series of questions the pupil is asked to express his degree of agreement on the subject 'reading and my future', σ_v means 'degree of importance' one considers reading has in relation to future possibilities.

Similarly, every question (item) has a 'degree of power' completely characterized by a parameter θ_i which is also a theoretical number to be estimated. The frame of reference is often illustrated like the figure. For the sake of convenience we note at the edge of frame the total score values T_v for each person and at the bottom the item totals.

Latent parameters / person	$1 \cdots i \cdots K$ $\downarrow\theta_1 \cdots \theta_i \cdots \theta_K$	$\leftarrow$ latent 'powers'
$1 \quad \sigma_1$		T_1
$\nu \quad \sigma_\nu$	{Responses}	Score totals $T_\nu = \Sigma x_{\nu i}$
$N \quad \sigma_N$		T_N
	Item totals	

3. Responses-data

The data obtained from a questionnaire like 'why I like to read' consist of categorical information. In fact, the questionnaire allows for four types of responses: (one agree with it ---) 'A lot', 'A little', 'Not much' and 'Not at all'. It is of great advantage for the statistical analysis that all items have been scaled equally, i.e. the same four answer categories have been used throughout the questionnaire. At least, the sub groups of items belonging to one specific subject, later called components or dimensions, should allways be scaled equally.

The name of response category is denoted **ncat**, and the four choices are just labeled $\mu = 1,2,3$ and 4. If we call the variable describing the response for person No. ν to item No. i for $X_{\nu i}$ we get, finally

(1) $X_{\nu i} = \text{ncat}(\mu)$ $\mu = 1,2,3,4$ $(4 = M)$

as notation for the outcome: 'response was μ for person No. ν to question (item) No. i'. The number of response categories, M is equal to four but in some of the statistical analyses response category No. 4 and No. 3 are collapsed into one category, since very few persons have used the extreme category: 'Not at all'.

Questions are those of the original questionnaire with references to numbers from the questionnaire. **Items** are those questions which are of interest in a sub group of items. Such sub groups are called **components** or **dimensions** with a term from the factor analytic tradition, although we are not going to perform this kind of analysis. We enumerate items simply by 1,2,3 ... to maximum number, e.g.

DIM6:	Question Nos.	22	33	36
	Item Nos.	1	2	3

4. The Rasch model

The statistical model used for the analysis of data is the so-called one dimensional version of the general Rasch model (Rasch, 1968). It calculates the probability of response ncat(μ) from person No. ν to item No. i as follows:

$$(2)\quad P(X_{\nu i} = \text{ncat}(\mu)) = \frac{e^{\varphi(\mu)(\theta_i + \sigma_\nu) + \varrho(\mu)}}{\gamma_{\nu i}}$$

There are 'many hidden reasons' for putting a mark on e.g. response category 'A little', which is **not** considered in **any** statistical analysis. The frame of reference with explicit reference to item power θ_i and personal 'degree' σ_ν sets up the limit for what is taken into account in this statistical analysis and what is not taken into account. The basic argument for applying a probabilistic description of responses as outcomes of an experiment is in fact the existence of these many hidden reasons. Reasons outside the range of what can be calculated if θ_i and σ_ν were known and therefore, must be treated as 'random influence'.

Besides the frame parameters θ_i and σ_ν the model contains $\varphi(\mu)$ and $\varrho(\mu)$, called **scoring function** (or scoring parameters) and **category function** (or category parameters). The scoring function plays the role of transforming the categorical observations ncat(μ) on to real numbers $\varphi(\mu)$ so that comparisons between persons can be carried out by means of simple scores. The majority of statistical analyses of questionnaire data avoid these problems by simply **assuming** (without proof) the scoring values 1,2,3 and 4 to be true.

The $\varrho(\mu)$ parameters are quantities reflecting the 'general use' of answer category ncat(μ). The $\gamma_{\nu i}$ parameter is a normalizing constant to assure that probabilities of ncat(1), ncat(2), ncat(3) and ncat(4) sum to unity.

5. Test of fit.

The one dimensional Rasch model (2) is a specialization of the general Rasch model for M response categories, which is the following:

$$(3)\quad P(X_{vi} = \text{ncat}(\mu)) = \frac{e^{\theta_i(\mu) + \sigma_v(\mu)}}{\gamma_{vi}}$$

In this model every person is characterized by a M-dimensional parameter $\sigma_v = (\sigma_v(1), \sigma_v(2), \sigma_v(3), \sigma_v(4))$ and likewise every question characterized by a M-dimensional parameter $\theta_i = (\theta_i(1), \theta_i(2), \theta_i(3), \theta_i(4))$. η_{vi} is again a normalizing constant, like γ_{vi}.

Test of fit for the one-dimensional model is based upon two main principles. First, a test of fit for the general model (3) is performed and if this model works it is tested whether its specialization (2) works also. Secondly, various consequences of accepting the one dimensional model are brought into light and tested against data. This twofold testing procedure has been selected because very general aspects concerning **measurement** are laid down in the general model while specific questions, like change in attitude towards 'why I like to read' over age, are precisely considered and evaluated more simplistic by the one-dimensional model.

Test[1)] of fit of the general model is based upon conditional distributions, in which we isolate the one set of parameters, e.g. all $\theta_i(\mu)$'s (item powers) denoted $((\theta_i(\mu))$. To obtain this distribution we need $\mathbf{r}_v = (a_v(1), a_v(2), a_v(3), a_v(4))$ which is for every pupil: $a_v(1)$ = number of responses given to category one over all items, $a_v(2)$ = number given to response category two --- etc.

The conditional distribution of all responses given these $\mathbf{r}_v$ - statistics, $(\mathbf{r}_v)$ is then the basis for estimation and test of the model:

$$L(((\theta_i(\mu))) \mid (\mathbf{r}_v)) = \text{Prob}\,(((X_{vi})) \mid (\mathbf{r}_v))$$

By means of this distribution we compare results over different age groups and over the two sex groups through ordinary likelihood ratio test statistics.

In fact, it can be shown (Andersen 1973) that

$$-2\log\,(L_1 \cdot L_2 \cdot \ldots \cdot L_q) \,/\, L_{tot} \sim \chi^2(df)$$

$$df = (q-1)(M-1)(K-1)$$

with q sub groups of total data giving rise to maximum values of likelihoods $L_1, \ldots, L_q$ to be compared with likelihood for total data L_{tot}. Note, that all L's are conditional distributions.

1. The following has interest for those interested in technical aspect of Rasch model analysis.

Numerical test of fit for the one-dimensional model can be brought about in two ways. Either we adopt likelihood ratio tests parallel to test of fit of the general model shown above or we evaluate one-dimensionality as a parametric hypothesis **within** the general model (Allerup 1985). The first procedure has been preferred in this work and it remains to note, that degrees of freedom for the χ^2-test is in this case df = (q-1)(K + 2M-5) since in the one-dimensional model we shall estimate one dimensional powers θ_i, scorings $\varphi(\mu)$ and category parameters $\varrho(\mu)$ (Allerup, 1985).

For the comparison of test results over different groupings of data with the one-dimensional model conditional probabilities of responses play a central role. Any statement based upon conditional probabilities **given** total score T_r is independent of individual parameter σ_r and is therefore, one side frame dependent only (dependent on item powers and scorings).

The probability of a specific response ncat(μ) given total score $T_r = t$ can be obtained as follows

$$P(X_{ri} = \text{ncat}(\mu) \mid T_r = t) = e^{\varphi(\mu)\theta_i + \varrho(\mu)} \frac{\Lambda^{(i)}_{t-\varphi(\mu)}}{\Lambda_t}$$

the Λ's entering this expression are quantities being quite cumbersome to calculate:

$$\Lambda_t = \sum_{\Omega} \prod_{(i)} e^{\varphi(\mu)\theta_i + \varrho(\mu_i)}$$

the summation is extended over the set Ω of **all** response combinations $\varphi(\mu_1),\ldots, \varphi(\mu_K)$ to the K items where total score $T_r = \Sigma x_{ri} = \Sigma\varphi(\mu_i) = t$ is fixed to t. $\Lambda^{(i)}_{t-\varphi(\mu)}$ is like Λ_t with item No. i omitted and total score t reduced to $t-\varphi(\mu_i) = t-\varphi(\mu)$ since ncat(μ) is the response to item No. i

It has furthermore been used to combine the conditional probabilities to something more simplistic, easy to overlook. Under the one-dimensional model X_{ri} is 'quantified' by scorings $\varphi(1),\ldots,\varphi(4)$ and it can - to some extend - be considered as an ordinary real random variable. In this sense it has proved usefull to include **conditional expected responses** $E(X_{ri} \mid T_r = t)$ into evaluation of fit. These expected values are simply

$$E(X_{ri} \mid T_r = t) = \sum_{(\mu)} \varphi(\mu)\, P(X_{ri} = \text{ncat}(\mu) \mid T_r = t)$$

where the probabilities are calculated above.

As a technical remark, we state the equation by which estimation of the latent σ_r-parameters takes place from an observed score value $T_r = t$. We equate **observed** T_r-score with **expected** T_r-score under the model:

$$T_r = E(T_r)$$

Since $T_r = \Sigma X_{ri}$, and the X_{ri}'s are now scored as $\varphi(1),\varphi(2),\varphi(3)$ and $\varphi(4)$ we then get

$$T_r = \Sigma E(X_{ri})$$

$$= \Sigma\Sigma\varphi(\mu)\ P(X_{ri} = \varphi(\mu))$$

$$= \Sigma\Sigma\varphi(\mu)\frac{e^{\varphi(\mu)(\theta_i + \sigma_r)\ +\ \varrho(\mu)}}{\gamma_{ri}}$$

in which estimated values of φ's, ϱ's and item parameters θ_i's are are assumed to be known. The θ's are estimated from conditional maximum likelihood equations - independent of σ_r's - and inserted here, we will not go into a discussion of conditional estimation of σ_r's by this occasion.

6. Interpretations of tests

An essential part of applying the Rasch models is the test of fit - or rather 'control of model'. This feature is going to draw upon attention in greater detail than met elsewhere in statistical analyses.

The reason for this can be sought in the origin of the statistical model. Factor analysis is, e.g. resting upon certain mathematical assumptions of normal distribution, various principles for rotation of the 'solution'... etc. **ending up** with the components or factor-dimensions, which are then to be interpreted by the user, the psychologist.

Rasch models go the other way round by explicitly adopting psychological concepts as a **starting point**. Given these conditions, the Rasch model is the necessary and sufficient solution to two equivalent requirements, both of which are familiar to daily, psychological work with questionnaire data (Rasch, 1968,1980)

- Persons must be compared in an objective way.
  Items must, likewise be objectively comparable (homogeneity)

- The many single-item responses in a dimension must be reduced to **one** score value.

Within the given frame of reference the first requirement means that a comparison of any two σ_v-values must be done in such a way that the result is independent of other σ_v-values. A simple demand which is not met in analyses of populations where e.g. **mean values** play a role.

The other requirement refers to a process which is usually overlooked as a problem. Most statistical analysis of questionnaire data takes it for granted to calculate a score value - for two response categories e.g. it means just counting the number of 'positive' answers. For more than two response categories (M), the reduction to scores means reduction to M-dimensional arrays of countings:

	response	μ-label
Item No. 1	'Not at all'	4
2	'A lot'	1
3	'A lot'	1
4	'Not much'	3
5	'A little'	2

Reduction to
4-dimensional score:

	total 1's	2
	total 2's	1
	total 3's	1
	total 4's	1
=	(2,1,1,1)	

or we write $a_{\nu.}$ = $(a_{\nu.}(1), a_{\nu.}(2), a_{\nu.}(3), a_{\nu.}(4))$
= (2,1,1,1)

with scoring function $\varphi(\mu)$:

$\varphi(1)$ =	1.0	or	4.0
$\varphi(2)$ =	0.8	or	3.4
$\varphi(3)$ =	0.2	or	1.6
$\varphi(4)$ =	0.0	or	1.0

We get the one-dimensional score:

$$2\ 1.0 + 1\ 0.8 + 1\ 0.2 + 1\ 0.0 = 3.0$$

The general Rasch model (3) is the solution to requirements for objective comparisons between M-dimensional σ_{ν}-parameters and the solution to the problem of modelling single-item responses if reduction to M-dimensional scores $a_{\nu.}$ is required.

Interpretation of tests therefore, means interpretation of the two equivalent conditions for the model. Or, which is the central point: rejection of the model means rejection of the two conditions, viz. it is **not** possible to perform objective comparisons and a calculated score does **not** contain all single-item information.

By attacking data from different angles, repeatedly using the model, we learn where the model fails - i.e. where the two conditions fail to work - and by this very fact, we in turn learn something new. An important angle is, for example, age specific changes towards 'why I like to read'. If the model works as a measuring instrument it must work 'equally well' on 8-year-old pupils and on 14-year-old pupils - like applying a simple ruler for measuring their heights. Of course we may come to different measures of an 8-year-old boy an a 14-year-old boy - this being expressed through different σ_{ν} values within the model - but the 'ruler' must be constant. We cannot allow for shifting units to occur on the ruler - at one occasion we measure in inches, at another cm's and yet another occasion some unknown units.

We think (hypothesize), that the questionnaire consists of fair, constant 'units', i.e. items. Test of model may reveal that it is not so by revealing age specific deviations from this assumption.

7. Some summary statistics

In appendix 1 the questionnaire which has been used is presented in its English version. Of course a translation into Danish was made, before the questions were given to Danish school children.

The questions touch various aspects of what a pupil feels about reading, and a first impression is, that the order in which the questions are given looks random. They are a.o. not grouped into sub-subjects in the lay-out of the questionnaire.

A total distribution of responses in the four response categories is given in the table below. This rather crude way of considering data tells only a little of the things we want to know about, but it may give some indication of the seriousness with which the pupils have completed the scheme. It gives, at the same time, a check on the 'spread' on response-category formulation. For instance, is 'A lot' far too weak an expression for ones agreement?

The first four columns refer to absolute countings - their totals being the fifth column (note that these totals are not constant due to missing responses). Relative distribution is given in the last four columns (their sum being equal to 1.00 - excluding missing observations).

Massive concentration (i.e. percentage from above 50%) of responses is accentuated by box-framing in the table. One can spot maximum concentration to the category agree 'A lot' on the questions (No. 10): 'I get enjoyment from reading' and (No. 13): 'Books tell good stories'.

Second to these maxima comes 'A lot'-agreement with (No. 8): 'It gets my mind off things and (No. 9): 'It makes me understand more'.

These questions are quite straightforward points about reading and one would generally, expect children to 'agree' with them. Technically it is satisfactory to observe, that all 100% are nowhere concentrated on one response category only.

It seems consistent with the general impression of the importance of religion to Danish school children, that they respond negatively to question (No. 6): 'Know more about my religion'. It is fair to say, that it is likewise expected that Danish school children are not very much occupied with Government wishes. Therefore, question (No. 47): 'The Government in my country wants me to be able to read' shows massive concentration on category (Agree) 'Not at all'. The majority of children seem to have understood the question 'correctly', in the sence that they are **not** answering a question concerning their benefit of learning to read.

The 50% concentration on response category (agree) 'A lot' with question (No. 22) 'Help me pass my examinations', demonstrates the consciousness about the importance of being able to read for advancing in the school grades.

In question (No. 33): 'People will praise me for being a good reader' quite many (50% + 25%) express their agreement as 'A lot' or 'A little'. In the sense that one should imagine the family gathered around the children, reading aloud, the father proudly applauding, the response to

this picture is surprising. However, in the Danish translation the result: 'being a good reader' is stressed at the cost of 'being praised', and the high percentages can be explained from this fact.

To a great extent, Danish children read, even if they are not forced to do so. This can be seen from question (No. 28): 'when I'm not forced to read', getting approx. 75% agreement 'A lot' or 'A little'.

Besides considering single questions where massive concentration is found on one or two answer categories one should also look at combinations of responses on selected questions. For instance, the four above mentioned straightforward questions (No. 8, 9, 10, 13) received many agree's 'A lot' or 'A little', but **some** children (1%, 3%, 4% and 1% respectively) did express their total disagreement with these questions. It is rather odd to imagine seriously responding children being negative towards **all** four questions, and it is satisfactory to find only one person (90% negative to all questions) in this group.

It is of the summary survey to make further checks on response patterns, i.e. looking for less probable ways of responding to two or more questions if one is really serious about answering. We have, e.g. looked for questions with some degree of repetition of the content. Such pairs of questions can be found as

- No. 8: 'gets my mind off things'
- No. 11: 'stops me from getting bored'

and

- No. 33: 'being a good reader' (cf. remarks above)
- No. 22: 'helps me pass examinations'

To demonstrate such pattern checks the tables in the following two figures show how responses to pair-questions are correlated. The majority of answers are given such that they correspond well, i.e. the diagonal have nearly all observations. As a curiosity we found among the six low outliers in the last table two eight-year-old girls giving exactly parallel responses to all questions, evidently attaching great importance to the **geometrical** layout of marks.

		'gets my mind off things' (No. 8)			
		alo	ali	nmu	na
'stops me from getting bored' (No. 11)	alo	107	91	59	68
	ali	58	135	120	106
	nmu	22	59	82	100
	na	20	41	65	102

alo = 'A lot', ali = 'A little', nmu = 'Not much' and na = 'Not at all'.

		'being a good reader' (No. 33)			
		alo	ali	nmu	na
'helps me'	alo	544	72	10	9
pass exa-	ali	67	201	42	7
minations'	nmu	10	31	105	29
(No. 22)	na	6	13	21	97

Apart from such peculiarities it is a general impression that the questionnaire did work technically well: Response category formulation offered a suitable range of 'values', the content of questions was well understood and a sufficient degree of seriousness was apparently present during the completion of the scheme - if we conclude on 'pattern checks' for questions expected to correlate much.

These calculations done on the total data set seem necessary before entering into a more detailled evaluation of the questionnaire and the data.

	ABSOLUTE NUMBERS					RELETIVE NUMBERS				
	A LOT 1	A LITTLE 2	NOT MUCH 3	NOT AT AL 4	TOTAL	A LOT 1	A LITTLE 2	NOT MUCH 3	NOT AT AL 4	
1:IS EXCITING	812	382	66	16	1276	0.636	0.299	0.052	0.013	
2:ANOTHER WORLD	787	366	92	33	1278	0.616	0.286	0.072	0.026	
3:HELPS ME SCHO	517	356	252	150	1275	0.405	0.279	0.198	0.118	
4:IT IS EASY.	336	403	276	251	1266	0.265	0.318	0.218	0.198	
5:INTERESTING.	659	424	134	57	1274	0.517	0.333	0.105	0.045	
6:KNOW RELIGION	122	173	244	688	1227	0.099	0.141	0.199	0.561	
7:HELPS STUDY	449	405	243	165	1262	0.356	0.321	0.193	0.131	
8:OFF THINGS	210	328	328	385	1251	0.168	0.262	0.262	0.308	
9:UNDERST MORE	537	410	244	78	1269	0.423	0.323	0.192	0.061	
10:ENJOYMENT	885	264	77	52	1278	0.692	0.207	0.060	0.041	
11:BEING BORED	338	431	266	234	1269	0.266	0.340	0.210	0.184	
12:PUTS THOUGHT	324	464	306	170	1264	0.256	0.367	0.242	0.134	
13:GOOD STORIES	866	323	62	18	1269	0.682	0.255	0.049	0.014	
14:GET GOOD JOB	398	356	312	198	1264	0.315	0.282	0.247	0.157	
15:PARENT THINK	336	236	270	428	1270	0.265	0.186	0.213	0.337	
16:FORGET WORRI	275	328	289	379	1271	0.216	0.258	0.227	0.298	
17:OTH.COUNTRIE	316	366	348	243	1273	0.248	0.288	0.273	0.191	
18:GOTO COLLEGE	400	337	251	262	1250	0.320	0.270	0.201	0.210	
19:I IN STORIES	276	397	313	288	1274	0.217	0.312	0.246	0.226	
20:SLEEP EASILY	400	288	245	347	1280	0.312	0.225	0.191	0.271	
21:RESPECT PEOP	267	391	369	243	1270	0.210	0.308	0.291	0.191	
22:HELP MY EXAM	646	320	176	137	1279	0.505	0.250	0.138	0.107	
23:I AM LONELY	480	481	169	145	1275	0.376	0.377	0.133	0.114	

	ABSOLUTE NUMBERS					RELETIVE NUMBERS				
	A LOT 1	A LITTLE 2	NOT MUCH 3	NOT AT AL 4	TOTAL	A LOT 1	A LITTLE 2	NOT MUCH 3	NOT AT AL 4	
24:LEARN NEW	453	461	259	103	1276	0.355	0.361	0.203	0.081	
25:MAKES HAPPY.	183	368	409	309	1269	0.144	0.290	0.322	0.243	
26:BETTER WORLD	176	345	394	345	1260	0.140	0.274	0.313	0.274	
27:NOTHING TV.	350	401	237	284	1272	0.275	0.315	0.186	0.223	
28:NOT FORCED	637	341	142	149	1269	0.502	0.269	0.112	0.117	
29:BOOK FRIEND.	405	425	245	201	1276	0.317	0.333	0.192	0.158	
30:ANOTHER TIME	301	378	339	250	1268	0.237	0.298	0.267	0.197	
31:WORK BETTER	316	341	344	271	1272	0.248	0.268	0.270	0.213	
32:WHAT RIGHT.	358	416	281	214	1269	0.282	0.328	0.221	0.169	
33:GOOD READER.	631	319	178	142	1270	0.497	0.251	0.140	0.112	
34:WAY TO LIVE.	174	320	364	401	1259	0.138	0.254	0.289	0.319	
35:IMPORT.PERS.	145	228	334	563	1270	0.114	0.180	0.263	0.443	
36:MORE WORDS.	524	368	250	135	1277	0.410	0.288	0.196	0.106	
37:LIVE BETTER	190	335	382	359	1266	0.150	0.265	0.302	0.284	
38:NICE ALONE.	535	412	173	152	1272	0.421	0.324	0.136	0.119	
39:HELP COUNTRY	228	262	324	442	1256	0.182	0.209	0.258	0.352	
40:KNOW WORLD	302	399	316	247	1264	0.239	0.316	0.250	0.195	
41:NICE TO THINI	387	459	299	126	1271	0.304	0.361	0.235	0.099	
42:IMPROV.WRITE	323	331	303	310	1267	0.255	0.261	0.239	0.245	
43:QUICK-SLOWLY	491	350	229	196	1266	0.388	0.276	0.181	0.155	
44:HELPS RELAX	477	422	211	162	1272	0.375	0.332	0.166	0.127	
45:TIME GO BY	410	400	259	197	1266	0.324	0.316	0.205	0.156	
46:TEACH.IMPORT	384	269	221	391	1265	0.304	0.213	0.175	0.309	
47:COUNT.WANTS	142	158	242	695	1237	0.115	0.128	0.196	0.562	
48:WORK BETTER	332	322	331	278	1263	0.263	0.255	0.262	0.220	
49:NOTH. TO DO	221	310	321	416	1268	0.174	0.244	0.253	0.328	
50:HELP MY COUNI	211	253	335	451	1250	0.169	0.202	0.268	0.361	

8. How to define components.

The fifty questions belong to various 'functions' or sub-subjects of reading. One could choose to analyse every one of the fifty questions, examine response patterns for girls, for boys - old and young - correlating the responses and try collecting the obtained multidimensional results in a few precise words. This is of course impossible.

On the other hand one would probably find conclusions drawn on the basis of one single score summed over all fifty questions to be too 'crude', too little reflecting the many reasons for a given response.

In between these two alternatives comes the traditional solution by cutting up the whole series of fifty questions into 'components' or 'dimensions' as we call them in the Danish data. The first problem is then how to define them.

One way, used by Greney and Neumann in their investigations - and in fact used nearly always in such data - is to base definitions of components on correlations between responses to the questions. Grouping of highly correlated questions, excluding questions with low intercorrelations, is essentially what a factor analysis does. Interpretation of 'components' or factors resulting from this type of statistical analysis is mainly of a mathematical nature and psychological arguments are always subsequent to mathematical arguments.

Definition of **components**

- + factor analysis

- ÷ psychologist

- ÷ educator

The factor analysis actually identified five significant factors, or components (see Greney and Neumann, 1984). These were

COMP1: Utilitarianism

COMP2: Enjoyment

COMP3: Moral

COMP4: Diversion

COMP5: Escapism

In our investigations we started by analysing the questions solely from psychological and educational view points. This led to a classification of the questions into six different 'subjects' or **dimensions**, as we call them. Note that this procedure was carried out by an expert on the main object: reading, no statistical means were involved at this stage. The Rasch analysis will go on from this point, taking these classifications as initial frame of references (six frames!). The further statistical analyses within each dimension will then, possibly, modify the content of dimension by

eliminating single items (like the factor analysis but on different grounds).

Definition of **dimensions**

- + Rasch analysis

- + psychologist

- + educator

The six dimensions come out to be

DIM1: Enjoyment

DIM2: Future

DIM3: Diversion

DIM4: My country

DIM5: Professional skill

DIM6: Improvement of reading

The items entering these dimensions are the following

DIM1: Nos:	1,2,4,10,13,19,25,28,29,38,41,44	**12 items.**
DIM2: Nos:	3,7,14,15,18,26,31,34,35,37,40,42,48	**13 items.**
DIM3: Nos:	8,11,16,21,27,45,49	**7 items.**
DIM4: Nos:	6,32,39,46,47,50	**6 items.**
DIM5: Nos:	5,9,12,17,20,24,30	**7 items.**
DIM6: Nos:	22,33,36	**3 items.**

For the identification of dimensions you are referred to appendix 1.

A close look at the two definitions of components and dimensions reveals that some DIM's are in good correspondence with some COMP's:

DIM2 ≈ COMP1
DIM1 ≈ COMP2
DIM3 ≈ COMP4

The stated equalities (≈) do not mean exact correspondence. It signifies generally that items met in components are **also** found by this primary definition of a dimension. It is to be remembered that the components are dealing with only 23 questions in total - the rest, 29 questions, were excluded during the factor analysis. The Rasch analysis will show whether we end up with reduced dimensions also, and whether the reduced DIM's still have items in common with COMP's.

One should keep in mind, that the aim of factor analysis is not the same as that of Rasch model analysis. Each of the two ways of analysing data gives qualifications to the questionnaire which is conceptually quite different (see above basic notation in Rasch Models).

Like COMP5, DIM6 has few items - although not equal by content. The dimension will not be considered by the Rasch analysis for technical reasons. Since Rasch analysis is concerned with the question of reducing single item information to one score value it looses some interest if we have only three items. Further, statistical tests of models become weak when only few items are considered.

9. What can be assessed by Rasch models?

The main objectives of our study is the following

- to see if the questionnaire can be used as an instrument for 'objective' measurements.
- to see if comparisons between response patterns from different pupils can be carried out by means of simple score values
- if, in fact so we have 'learned something' to report concerning items eliminated during this process. On the remaining items (reduced dimension) we carry out comparisons by means of total score values.

In the following sections we shall mix theoretical observations on these problems with empirical results from the questionnaire data. The first point for discussion will be how scoring of response categories affects the understanding of data. It will be a red thread for Rasch model analyses to look for contrasts in data and after that signify and interpret the conceptual content of 'misfits', eventually obtained, finally reporting it as one result of analysis. In fact, we shall start with the situation in which non-constant scorings (i.e. varying φ-values in the model) may bring into light certain characteristics about data, or in turn certain characteristics about the people behind the data, i.e. the pupils.

As mentioned earlier we put it off till later to draw consequences of the Rasch model when it is actually fitted to data.

10. The scoring system.

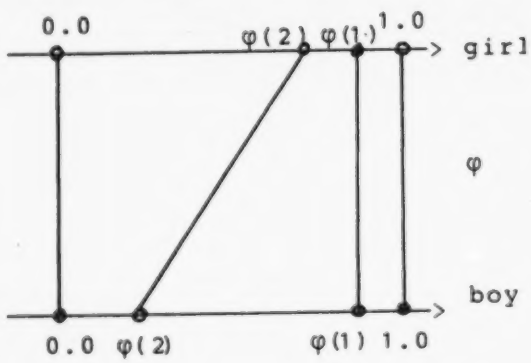
It shall be noted as a first remark, that the usual total score value has no meaning, unless every response category has been 'scored' properly, like for instance

'A lot'	≡ 4.0	= $\varphi(1)$
'A little'	≡ 3.9	= $\varphi(2)$
'Not much'	≡ 2.1	= $\varphi(3)$
'Not at all'	≡ 1.0	= $\varphi(4)$

This example shows how the response categories can be scored in a way, which is different from the simple scorings 1,2,3,4 - often taken as valid scorings **without** even considering that it poses a problem for statistical evaluation. In some dimensions we may find one scoring system $\varphi(1)$, $\varphi(2)$, $\varphi(3)$ and $\varphi(4)$, in another dimension these values may be quite different. This is in accordance with the analysis under the Rasch model since we are dealing with five (six) frame of references, each one estimating 'its own' set of parameters in the model.

It is not allowed, however, to have different scorings **within** one and the **same frame** of reference it not allowed, cf. the remarks above on shifting 'dividing points' on a ruler. If we find such tendencies e.g. when comparing boys and girls what does that mean?

	♂ (A)	♀ (B)
$\varphi(1)$	1.0	1.0
$\varphi(2)$	0.9	0.9
$\varphi(3)$	0.8	0.4
$\varphi(4)$	0.0	0.0



We imagine that the only difference between the two scoring systems is the way in which response category No. 3, - 'Not much' - is scored (from technical reasons response categories No. 1 and No. 4 are constrained to the values 1.0 and 0.0).

In terms of the Rasch model we obtain a series of consequences from the above situation, some of them suitable for **interpretation** of such scoring system differences, some of them specifically derived as **test statistics** since it is **not** acceptable that boys have one set of scorings, girls another. To this end it is necessary to evaluate whether the (estimated) difference 0.8-0.4 is statistically significant or not. This is done as part of test of fit for the one-dimensional model as a whole. We start looking for consequences of scoring system A and B by reminding you that calculation of the total scores is at all interesting because of the latent parameter σ_r.

In fact, in the model

$$P(X_{vi} = \text{ncat}(\mu)) = \frac{e^{\varphi(\mu)(\theta_i + \sigma_v) + \varrho(\mu)}}{\gamma_{vi}}$$

we estimate the latent degree of agreement σ_v by means of total score T_v (sufficient statistic)

$$T_v = \Sigma x_{vi} = \varphi(\mu_{vi}) + \ldots + \varphi(\mu_{vK})$$

The outcomes x_{vi} of the variables X_{vi} are just μ-category-names

$\mu_{vi},\ldots,\mu_{vK}$ whuich the pupil has chosen as responses to K items.

As an example we look at the 7 items in DIM5 and select a specific pupil's response pattern:

Responses: ' (3) 4 2 4 2 2 2'

↳ response to item No. 1 is affected by system A or B.

The scoring systems A and B will then result in the following two score values

A : $T_v = (0.8) + 0.0 + 0.9 + 0.0 + 0.9 + 0.9 + 0.9$

$= 4.4$

B : $T_v = (0.4) + 0.0 + 0.9 + 0.0 + 0.9 + 0.9 + 0.9$

$= 4.0$

Though score values $T_v = 4.4$ and $T_v = 4.0$ are 'basic' values for the analysis they are, in fact providing the **initial stage** only for the calculation of the parameter of 'real' interest, i.e. σ_v.[1] We see that the more '3's a response pattern contains the more different the scores T_v under A and B wil be.

1. Relation between 'raw' score values $\hat{T}_v$ and estimated, latent parameters σ_v needs not be a simple linear relationship. It has consequenses for evaluating the magnitude of differences - whether one does it on the score-scale or on the latent scale.

This gives rise to one important key for interpretation of differing scoring systems. In fact it is clear, that with scoring system B we **discriminate** better between response categories '3' and '2' than is done due to A scorings. This in turn affects total scores T_r in such away that 'relatively disagreeing' boys (i.e. with many 3's in the response pattern) get higher T_r-scores than their girl companions with the same response pattern.

For the value of individual parameter σ_r, the significance of differering scoring systems is more complex. We shall mention, briefly that in case of these seven items, estimated σ_r's under scoring system A, $\sigma_r(A)$ and $\sigma_r(B)$'s under scoring system B show the following characteristics:

Responses	T_r-score	$\hat{\sigma}_r$
'3 4 2 4 2 2 2' →	A: 4.4	-1.42
	B: 4.0	-0.90

	$\hat{\sigma}_r$	T_r-score
A	-5.23	1.0
B	-5.64	
A	-4.47	1.5
B	-4.49	
A	-3.85	2.0
B	-3.61	
A	-0.74	5.0
B	0.48	

This shows that we have

$$\sigma_r(A) < \sigma_r(B) \text{ when } T_r > 1.50$$

$$\sigma_r(A) > \sigma_r(B) \text{ when } T_r < 1.50$$

In other words, for some T_r-values the estimated (latent) level of agreement, σ_r will be lowered when shifting from scoring system A to B; for other T_r-values we observe the opposite effect.

So much for interpretational aspects of shifting scoring systems. We now turn for a while to consequences that have effects on the **validation** of the model - and by that we mean evaluation of conditional statements.

For instance, we may evaluate the probability of responding 'Not much' to item No. 1 in our example calculated for a given score level. This procedure leads to statements which are general with respect to the population used, i.e. the statements are independent of the σ_i-parameters.

response category No. 3	item No. 1 in DIM5		
T_i-score	A	B	C
0.8	0.58	0.56	0.56
1.8	0.13	0.23	0.33
2.6	0.32	0.29	0.29
3.4	0.45	0.31	0.09
4.2	0.54	0.07	0.07
5.0	0.60	0.08	0.05
5.8	0.37	0.02	0.02

The content of the table is the chance or probability of getting 'Not much' as response to the first item of DIM5 (question No. 2) in the questionnaire: 'I read because it is interesting'. We calculate this chance from various agreement levels, i.e. various T_i-score levels. In other words, we calculate the conditional probability

$$P(X_{ni} = \text{ncat}(3) \mid T_i = t)$$

$$\text{for } t = 0.8,\ldots,5.8.$$

The last column C is shows what was actually estimated from data - we found the φ's to be 1.0, 0.7, 0.4 and 0.0 based on these seven items in DIM5. Note, that A and B systems were simulated values.

Several observations can be read off from the table. In the C-column the values decrease systematically[2] with increasing T_i-score. We are considering an answer category that is 'negative' of content and with increasing T_i-scores we focus on pupils with many '1's, i.e. 'agree's in their response pattern. Therefore, the chance of being 'negative' with one item when you are generally very much 'positive' (high T_i value) actually, must be low, the more 'positive' you are generally, the lower.

In column A the values attain a minimum in the middle of T_i-range and

2. Because of influence of the numerical values of the φ's on the possible T_{ni}-values this should be understood within certain groupings of T_i-values.

maximal values near the ends of T_r-range. This points to the fact that the A-scorings (φ: 1.0, 0.9, 0.8, 0.0) **do not** work.

Tables like this one which serves as basis for judgments of similarity between e.g. boys and girls. It is then up to statistical evaluation to decide whether or not estimated differences obtained from two scoring systems, like those observed between column B and C, for instance, are statistically significant.

We end the general discussion about analyses regarding scoring systems by noting that actual statistical analyses of scoring systems in practical work run parallel to analyses of homogeneity described later.

11. DIMENSION 1: Enjoyment

The statistical analyses of this dimension were rather confusing. 12 items are build into the dimension in the first place, but initial Rasch-model-analyses soon revealed that some items were sex-specific, some other items proved to be age-specific, so at the end it is still a problem whether or not a reduced DIM1 will emerge from the analyses.

It was, however, clear from the first step of analysis in which all twelve items were considered, that large differences in the scoring systems were present.

12 items

	$\varphi(1)$	$\varphi(2)$	$\varphi(3)$	$\varphi(4)$
Total	1.00	0.57	0.15	0.00
Boys	1.00	0.64	0.27	0.00
Girls	1.00	0.51	0.04	0.00
Age 8	1.00	0.55	0.21	0.00
Age 11	1.00	0.54	0.11	0.00
Age 14	1.00	0.68	0.34	0.00

The two main grouping criteria are age of year and sex. Deviations between model and data will be looked for especially in these directions.

This does not, by any means, exclude other grouping criteria. On the contrary, one should allways carry out supplementary checks, and we used school number as an arbitrary selection factor. The results of these checks of fit will, however, not be reported here.

We view the observed differences in terms of **discriminating** power between response categories (see interpretation of φ's above) and we see a marked sex-dependency in the way in which boys and girls conceive 'Not much'. Apparently, boys distinguish between 'Not much' and 'Not at all' more distinctly than girls - in this context! Now, analyses of homogeneity presented later on point especially to items Nos. 2, 4 and 10 as reasons for lack of homogeneity.

Item 2 (question 2): 'it makes me feel like I'm going to another world...'

Item 4 (question 10): 'I get enjoyment from reading'

Item 10 (question 38): 'it is a nice thing to do when I am alone'.

We may therefore, cautiously conclude, that different discriminations between the two negative response categories are related mainly to these three questions.

The sex analyses can be followed by similar conclusions for the age groupings. In fact, a distinct trend towards better discrimination

between the two negative response categories with increasing age level between the two negative response categories is observed. Both sex and age groupings have relatively stable $\varphi(2)$'s, i.e. rather consistent 'understanding' of the two positive response categories.

We shall not go into detail with subsequent searches for sub-dimensions of DIM1 as regards the scoring system. These analyses demonstrate conclusions taken from the first step in which all twelve items of DIM1 were considered, and it becomes clear that DIM1: 'Enjoyment' in its 12-item form displays difficulties especially for the scorings of negative response categories.

12. DIMENSION 2: Future

Dimension number 2 contains in its primary stage 13 items. It is by that the dimension having most questions. The first step of analysis in the series of steps searching for a homogeneous, reduced dimension with independent scoring system (depending on response categories only) revealed the following estimated scoring system:

13 items

	$\varphi(1)$	$\varphi(2)$	$\varphi(3)$	$\varphi(4)$
Total	1.00	0.63	0.31	0.00
Boys	1.00	0.59	0.29	0.00
Girls	1.00	0.68	0.33	0.00
Age 8	1.00	0.58	0.29	0.00
Age 11	1.00	0.66	0.33	0.00
Age 14	1.00	0.68	0.37	0.00

An immediate impression from these figures is, that sex grouping has caused minor differences in score values for response category number 2 and that age grouping points to systematically varying score values for both response category 2 and 3.

These considerations are reflected in the numerical tests of fit of both the general Rasch model and its one-dimensional specialization. In fact, first-step tests (see below in chapter on homogeneity) indicate strong rejection of both versions of the Rasch model in case of grouping by age variable but show less distinct 'misfit' when the results from the two sex groups are taken together.

The tests show bad adjustment of Rasch model to data especially for

items Nos. 1, 2, 4 and 5:

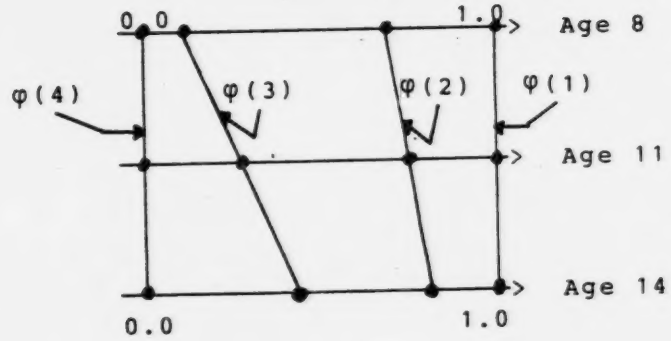
Item 1 (question 3): 'it helps me at school'.

Item 2 (question 7): 'it helps me with subjects I will study later on'.

Item 4 (question 15): 'my parents think it is important'.

Item 5 (question 18): 'I read so I will learn how to respect other people'.

These four questions are the main factor behind the lack of fit in the first step of analysis, and since poor fit is observed especially when we combine results from the three age-groups we shall concentrate on interpreting these test results. In fact, we may refer the systematics found in scoring systems over age groups to these four deviating items.



In case of maximal discrimination between the four response categories one should have 'equidistant' spacing between the scorings: $\varphi(1)$, $\varphi(2)$, $\varphi(3)$ and $\varphi(4)$. This brings us to the values 1.00, 0.67, 0.33 and 0.00 which is often called 'natural scoring'. The shift of scoring system from Age 8 to Age 14 shown in the figure suggests an interpretation like the following.

Lower age groups distinguish less distincly between the three response categories 'A little', 'Not much' and 'Not at all' than higher age groups do. Especially with these four items young pupils are liable to classify answers 'black' or 'white', 'no' or 'yes'.

13. DIMENSION 3: Diversion

This dimension contains seven items and the first step in search for homogeneous items was surprising. All seven items passed the test of fit, (χ^2 = 18.7 on df = 18) when the grouping variable was sex and scorings of response categories valid for both sexes could be estimated as follows:

$\varphi(1) = 1.00$
$\varphi(2) = 0.70$
$\varphi(3) = 0.28$
$\varphi(4) = 0.00$

A slight tendency towards sex-bounded scorings - although not significant - was found, the specification of which is in agreement with results founded in DIM1 (girls having lower discrimination between 'Not much' and 'Not at all' than boys).

The analysis of fit in which grouping variable is age level showed, however, strong, significant deviations from the model - applied to the general version (3) as well as its one dimensional specialization (2).

Estimated scorings for each age group was found to be

Seven items

	$\varphi(1)$	$\varphi(2)$	$\varphi(3)$	$\varphi(4)$
Age 8	1.00	0.78	0.40	0.00
Age 11	1.00	0.58	0.10	0.00
Age 14	1.00	0.71	0.28	0.00

The drop of $\varphi(3)$ value for Age 11 is clear and it persists in other checks of these seven items. This oddity is not easy to interpret since its immediate explanation should be: **low** discrimination between response categories 'Not much' and 'Not at all' for Age 11 but substantially **higher** discrimination between the same two response categories for eight-year old **and** fourteen-year old pupils. The analysis points to the following three items of DIM3 as the main factors behind the lack of fit:

Item 3 (question 16): 'it helps me forget worries'.

Item 5 (question 27): 'I read when there is nothing exciting on TV'.

Item 4 (question 21): 'it helps me to go to sleep easily'.

Changes in scoring systems - especially for the third response category are therefore, to be seen in connexion with these three questions.

When we omit the three items, homogeneity is improved (see later) and the scorings become stable over sex and age groupings. In fact, we end up with the following estimated values from the total data.

$$\varphi(1) = 1.00$$
$$\varphi(2) = 0.69$$
$$\varphi(3) = 0.32$$
$$\varphi(4) = 0.00$$

These scoring values can as well be presented as the following numerical values

$$\varphi(1) = 4.00$$
$$\varphi(2) = 3.07$$
$$\varphi(3) = 1.96$$
$$\varphi(4) = 1.00$$

From the transformed scorings we see that they are likely to be tested as 'natural scorings', i.e. $\varphi(1) = 4$, $\varphi(2) = 3$, $\varphi(3) = 2$, $\varphi(1) = 1$ which are the values normally used without further comments. As said it is to be preferred that the scorings are 'natural', since it simplifies calculations and signifies that the response categories are equally 'spaced'.

14. DIMENSION 4: My country.

There are six items in this dimension. First step of analysis clearly revealed problems of obtaining homogeneity and pointed to two (or may be three) items carrying the lack of fit, especially in the process of fitting the dimensional model in which the scoring system becomes crucial.

Item 2 (question 32): 'it will help me know what is right and what is wrong'.

Item 4 (question 46): 'it will help me to learn more words'

Item 1 (question 6): 'I read to know more about my country'.

Regarding the scoring system it was not, however, to say that estimated values showed great differences. Neither within sex groups nor within age groups. In fact, we found the following estimated scoring systems when all six items of DIM4 were considered:

Six items

	$\varphi(1)$	$\varphi(2)$	$\varphi(3)$	$\varphi(4)$
Total	1.00	0.77	0.43	0.00
Boys	1.00	0.76	0.39	0.00
Girls	1.00	0.79	0.45	0.00
Age 8	1.00	0.69	0.37	0.00
Age 11	1.00	0.67	0.36	0.00
Age 14	1.00	0.80	0.45	0.00

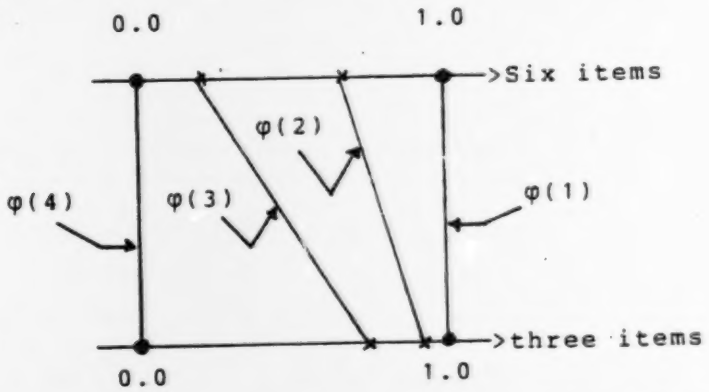
On the other hand, if the three items are omitted from the dimension we obtain the following estimates

Three items

	$\varphi(1)$	$\varphi(2)$	$\varphi(3)$	$\varphi(4)$
Total	1.00	0.87	0.63	0.00
Boys	1.00	0.88	0.58	0.00
Girls	1.00	0.85	0.67	0.00
Age 8	1.00	0.94	0.59	0.00
Age 11	1.00	0.74	0.66	0.00
Age 14	1.00	0.93	0.59	0.00

Now, with three items we managed to get satisfactory fits both to the general Rasch model (3) and the one dimensional version (2) (see the chapter on homogeneity). The displayed variation among estimated scoring values $\varphi(2)$ and $\varphi(3)$ is therefore, **not** significant i.e. the variation can be considered as 'random' variation.

Note the difference between the first scoring system with all six items and the last one accompanying the reduced set of items. A general **shift** of both $\varphi(2)$ and $\varphi(3)$ values can be observed when we exclude the three items:



It seems that the influence of the three items **as regards the scoring system** has been to 'maintain' four response categories rather than three, which is likely to be the situation[1)] with the reduced dimension consisting of three remaining items. These are the following

Item 3 (question 39):	'it will help my country if I learn things'.
Item 5 (question 47):	'the government in my country wants me to be able to read'.
Item 6 (question 50):	'I read so I will know how to help my country'.

Interpretation of estimated differences in scoring systems for six-item and reduced three-item could be, that the last three items are more suitable to decide upon using **three** rather than **four** response categories?

1. Since $\varphi(2)$ values are close to 1.0 discrimination between response categories Nos. 1 and 2 is weak.

15. DIMENSION 5: Professional skill.

This dimension is covered by seven items. First step in the search for homogeneity resulted in rather stable estimates of the scoring system both over grouping variables age level and sex. In fact, we found

Seven items

	$\varphi(1)$	$\varphi(2)$	$\varphi(3)$	$\varphi(4)$
Total	1.00	0.67	0.31	0.00
Boys	1.00	0.66	0.25	0.00
Girls	1.00	0.70	0.38	0.00
Age 8	1.00	0.63	0.33	0.00
Age 11	1.00	0.68	0.33	0.00
Age 14	1.00	0.69	0.33	0.00

Especially over age groupings the scoring system seems stable. However, the homogeneity analysis (see below) pointed clearly to one item, No. 1 being responsible for the fact, that all seven items cannot be used as a measuring instrument:

Item 1 (question 5): 'it is interesting'.

There may be good reasons for a 'misfit' of the general Rasch model - and of the one dimensional version as well - with item No. 1 as a member of DIM5 (see below on homogeneity), and it is therefore, reasonable to remove this item in the next step of analysis.

Considering the remaining six items we estimate, again the scoring system and from the estimates it becomes clear that only small differences in scoring values over age and sex groups can be detected. This is confirmed by the numerical fact that reduced DIM5 gives rise to satisfactory statistical fits of the general Rasch model **and** the one-dimensional version.

The scoring system for this six-item DIM5 and its transformation to the usual notation (which is exactly the same) is given as follows:

	Estimated	transformed
$\varphi(1)$	1.00	4.00
$\varphi(2)$	0.68	3.04
$\varphi(3)$	0.40	2.20
$\varphi(4)$	0.00	1.00

This scoring system is close to 'natural' scoring system: (1,2,3,4) like the values found for DIM3. Hence, the verbal formulation of response categories have led to satisfactory 'spacing' of response categories in these items.

As said earlier, DIM6 contains only three items in its first version. For

the time being we save our effects to dimensions with more items and we will report elsewhere how to deal especially with dimensions having such a small number of questions. Besides statistical problems of great uncertainty in parameter estimation and verification of models it is doubtfull whether a calculation of one score value from e.g. three single item responses is at all worthwhile doing.

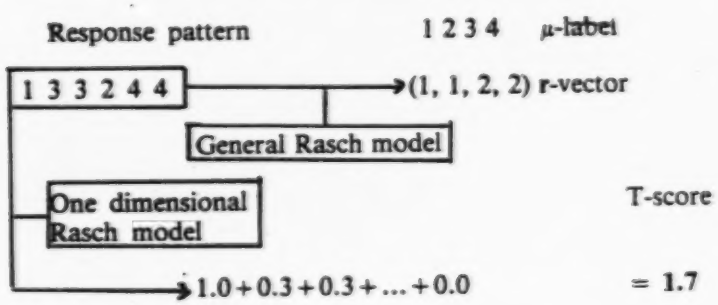
16. Analyses of homogeneity.

In this chapter we shall go into some detail with analyses concerning the search for homogeneous items. Those were defined as sub sets of items fitting the Rasch model - it being the general Rasch model (3) or the one dimensional specialization (2) above. In the first case we talk about four-dimensional measurements, in the latter we apply simple scorings of answer categories using the 'sum' (total score value) as a key for comparisons, and we talk about one-dimensional measurements; i.e. we talk about 'more' or 'less', 'up' or 'down'.

two-dimensional measurements one-dimensional measurements

Since we cannot draw four dimensional graphs we illustrate in two dimensions the basic idea of measerement in the general model. This is done in the first figure where 'direction' is a matter of two coordinates. As stated earlier the general Rasch model assumes the series of single item responses in a DIM to be reduced to a four-dimensional array of information:

1 = 'A lot'	$\varphi(1) = 1.0$
2 = 'A little'	$\varphi(2) = 0.8$
3 = 'Not much'	$\varphi(3) = 0.3$
4 = 'Not at all'	$\varphi(4) = 0.0$



In case of one-dimensional Rasch model we must estimate the scoring system and by that we obtain **one** value, 1.7 e.g. as total score value instead of **r**=(1,1,2,2) which is the four-dimensional total score in the general Rasch model.

The search for a sub set of homogeneous items starts by first testing the

general Rasch model against all items. This is carried out as a series of fits by means of various subdivisions of data. Since age level and sex will expectedly cause 'contrasts' in data, it is vital for the definition of homogeneity to use **explicitly** these basic variables in the testing procedures.

For each such variable both the general Rasch model and the one dimensional version is fitted and it is then evaluated by numerical methods and by graphical methods whether or not the model is an acceptable description of data. Numerical test statistics are likelihood ratio statistics (see Test of fit) and graphical evaluation consists of comparisons of comparing estimated item parameters, i.e. estimated 'powers' of items. If the Rasch model fits data, **all** subgroups of data must have a common set of item parameters and since the 'measurement' property of the DIM in question is important over age and sex we use exactly those variables to define our main groupings.[1)]

One of the key points for the Rasch models is **conditional inference**. By that we mean information collected as information at a **given, known** score level. We study response patterns **given** that we know the score total of that pattern. It is the possible 'variation' of a response pattern for fixed total score which is of interest.

Basicly we ask the question: What is the chance of responding 'a lot' to item No. 1 if we know the person's total score to be equal to some value? We can put this question within the Rasch models in contrast to e.g. factor analysis models because we have 'parametrisized' the individual response behaviour, i.e. we have allocated a parameter σ_v to each person. This parameter is estimated by the total score and since the score is sufficient (see basic notation in Rasch models) the above type of questions can be evaluated **independently** of person parameters. We have a.o. arrived at 'conditionally calculated chances' only depending on item parameters (and the scoring system) which can then be compared over different sub groups to see if the calculated chances are really independent of sub groups, i.e. are constant.

Differences found in sub groups by conditionally calculated chances or probabilities are 'summed up' in estimated item parameters (powers) and that is the reason for graphical analyses of item powers in the sequel.

1. As a control other kinds of groupings are advisable, e.g. random division of data.

17. The search for homogeneity.

We want to illustrate the principles outlined above for DIM5: Professional skill, in which case analysis showed to be simple.

The questions of DIM5 are

Item 1 (question 5):	'it is interesting'
Item 2 (question 9):	'it makes me understand more'
Item 3 (question 12):	'it puts thoughts in my mind'.
Item 4 (question 17):	'I read so I can learn how people live in other countries'
Item 5 (question 20):	'I read so I will learn how to respect other people'
Item 6 (question 24):	'It helps me learn something new'
Item 7 (question 30):	'I read so I can imagine what it's like at another time in history'.

We estimate the seven item powers of the general model and proceed to the one-dimensional model immediately.[1] Estimates of the scoring system are listed together with the results of likelihood ratio test statistics in the table below. All calculations are carried out on a computer where two 'packages' for Rasch model analysis are prepared to give all kinds of estimation and test results (see Allerup 1977, 1985).

one-dimensional Item powers in DIM5

Item	1	2	3	4	5	6	7
Total	2.62	1.14	-0.50	-1.17	-1.46	0.65	-1.28
Boys	2.37	1.05	-0.43	-1.11	-1.14	0.52	-1.30
Girls	2.98	1.22	-0.61	-1.27	-1.80	0.75	-1.26
Age 8	1.74	1.40	-0.63	-0.67	-1.31	0.52	-1.03
Age 11	2.42	1.20	-0.39	-1.14	-1.58	0.59	-1.11
Age 14	4.11	0.90	-0.53	-1.90	-1.56	0.90	-1.92

1. It will require much space to give the four-dimensional parameters for all sub-groups.

χ^2-fractiles

	general Rasch model	one dimensional Rasch model	Scoring system			
Total	-	-	1.00	0.67	0.31	0.00
Boys	} 94%	} 89.3%	1.00	0.70	0.38	0.00
Girls			1.00	0.70	0.38	0.00
Age 8			1.00	0.63	0.33	0.00
Age 11	} 100%	} 100%	1.00	0.68	0.33	0.00
Age 14			1.00	0.69	0.33	0.00

It is seen from χ^2-statistics in the tables, that both general model and one-dimensional version must be rejected. It is furthermore clear that rejection is bound up with unacceptable differences between the three age groups. In fact, the χ^2-fractiles 100% indicate a significance $p<0.0001$ for test of hypothesis: The three age groups have **one** set of item powers in common.

The items do **not** constitute a homogeneous set of items - we are **not** able to carry out objective measurements by means of these items - we are **not** in a position to compare responses between two pupils by means of total score values --- etc., lining up the consequences of rejecting the model. We have already taken the consequence of this rejection as concerns the scoring system and we shall now see how rejection is interpreted in terms of homogeneity.

From estimated item powers - assuming the one dimensional Rasch model to be true - we immediately get that item No. 1 in particular has greatly varying values over the age groups. Suspicions are therefore, directed to this item for causing lack of fit. However, this should be 'confirmed' by parallel judgement by the psychologist who is 'responsible' for the formulation of the question; why has it at all been included in the questionnaire? Are there troubles especially with the content of this question in relation to other questions of the same DIM (context problems)? -- etc.

In this case it is not surprising that the power increases with increasing age level. It means that the **relative importance** attached to this question (in relation to the other questions of DIM 5) is increasing with increasing age level. While the other questions of DIM5 have fairly precise references to the assertions stated, item No. 1 has a very broad, abstract formulation. Could it be, that 'interest in reading' is connected with the **amount** of books read so far?

We elaborate a little on the effect of varying power estimates for item No. 1. It can be illustrated how increasing item power for this item with increasing age level affects chance of responding in different answer categories. We base the calculations on parameters from total data for illustrative purposes.

DIM5 Seven items

		Item No.1	Item No.7
Total score 1.9:	ncat(1)	0.14	0.01
	ncat(2)	0.30	0.04
	ncat(3)	0.46	0.52
	ncat(4)	0.10	0.43
		1.00	1.00
Total score 6.0:	ncat(1)	0.91	0.59
	ncat(2)	0.07	0.17
	ncat(3)	0.02	0.23
	ncat(4)	0.00	0.01
		1.00	1.00

The table displays conditional probabilities of responding to each of the four response categories - conditional upon the 'knowledge', that total score is fixed to t = 1.9, 6.0. This probability is in the mathematical form (see Test of fit):

$$P(X_{pi} = \text{ncat}(\mu) \mid T_p = t) = e^{\varphi(\mu)\,\theta_i + \varrho(\mu)} \frac{\Lambda_{t-\varphi(\mu)}}{\Lambda_t}$$

The item power for item No. 1 is (see first table of this chapter) $\theta_1 = 2.62$ and for item No. 7 it is $\theta_7 = -1.28$. The ϱ-parameters have been estimated (Allerup, 1985) to 1.0, 1.7, 1.3 and 0.0.

It is quite clear from the calculations table that in case of item No. 1 all pupils have **great** probability of answering the first response category ncat(1) - 91% and 14% - while they have less probability giving the same response to item No. 7 - 59% and 1%.

The general effect of increasing item power is therefore, that agreement 'A lot' is more likely to occur for items with great item powers than for items with small item powers (fixed scoring system!).

It is, furthermore, clear from the table, how the general level of agreement, expressed as total score t affects probabilities of responding the four categories. Low-scoring persons (t = 1.9) are **not** responding 'A lot' to item No. 1 (14%), they are most likely to respond 'Not much' (46%). On items with less power, item No. 7 with $\theta_7 = -1.28$ such persons are yet more likely to respond 'Not much' (52%).

This is the background for the interpretations made above, where increasing item No. 1 -power over age levels were explained as increasing relative importance, attached to this question.

We end the discussion on homogeneous items in DIM5 by presenting 'the solution', i.e. the estimates for the remaining six items.

Item powers :	$\theta_2 = 1.78$	
	$\theta_3 = -0.09$	
	$\theta_4 = -1.11$	
	$\theta_5 = 1.20$	
	$\theta_6 = -0.95$	
Scoring system	$\varphi(1) = 1.00$	$\varrho(1) = 1.0$
	$\varphi(2) = 0.70$	$\varrho(2) = 1.7$
(and category	$\varphi(3) = 0.38$	$\varrho(3) = 1.4$
parameters)	$\varphi(4) = 0.00$	$\varrho(4) = 0.0$

These parameters will enter the calculations later on, when we compare pupils by means of simple 'rulers', i.e. six reduced, homogeneous dimensions.

18. Twisting items

Very often first step analyses point to more than one item being responsible for lack of fit. In DIM3, using all seven items it was surprising to see no significant differences if we grouped by sex variable. If age level was used, however, strong rejections of both the general Rasch model and its one-dimensional version were obtained.

The two heavily age-dependent questions are

Item 3 (question 16): 'it helps me forget about my worries'

Item 5 (question 27): 'I read when there is nothing exciting on TV'.

First we take a look at the distribution of answers given to these two questions:

	Item No. 3				Item No. 5			
ncat(μ)	1	2	3	4	1	2	3	4
Age 8	(134)	84	56	71	(114)	81	56	94
Age 11	84	(137)	90	121	126	(133)	81	92
Age 14	31	85	132	(165)	87	(165)	84	79

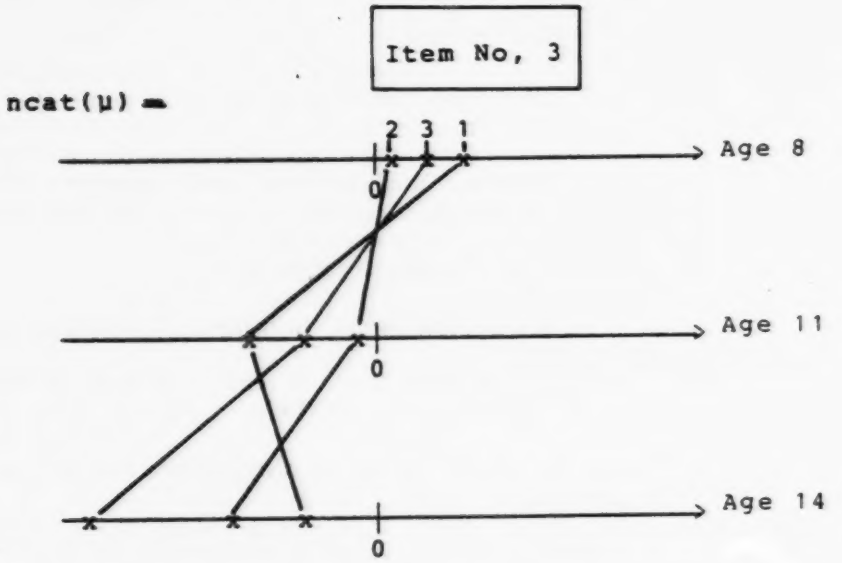
Distribution of responses (number of pupils)

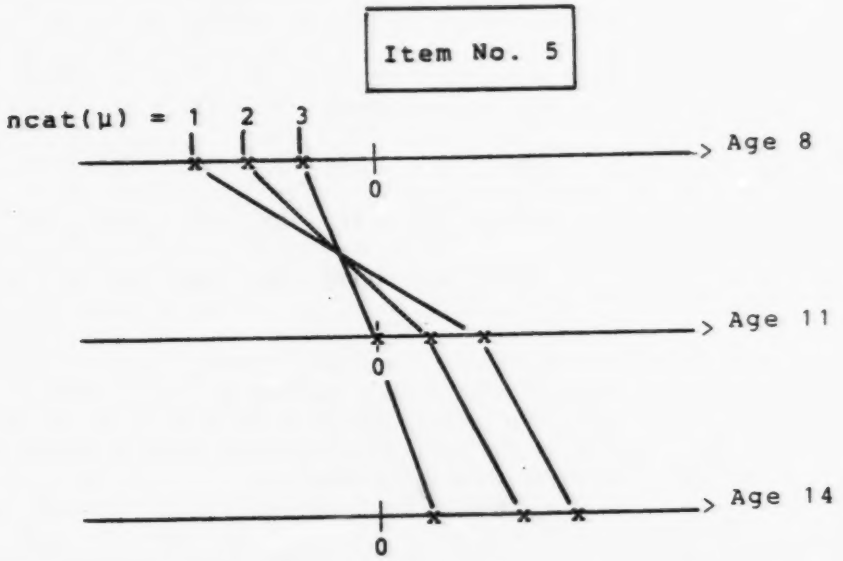
The distribution of responses evidently shows some systematics (incircled maxima) with shifting level of age. This **could** be, one might think, because a general change in the level of agreement takes place with shifting age level. The parameter estimates under the general model, however, show that it is **not** so:

	Item No. 3				Item No.5			
ncat(μ)	1	2	3	4	1	2	3	4
Age 8	0.24	0.18	0.19	0.00	-0.36	-0.26	-0.16	0.00
Age 11	-0.49	-0.10	-0.21	0.00	0.40	0.26	0.00	0.00
Age 14	-0.47	-1.14	-1.44	0.00	1.25	0.87	0.10	0.00

Four-dimensional item powers $\theta_i(\mu)$ for the general Rasch model are displayed in the table (parameter values for response category four are constrained to 0.00).

The structure of the table with four-dimensional item powers is easier to overlook graphically:





If the general Rasch model were a true description of data with all seven items present in the dimension, we should find constant ncat(1) - parameters, ncat(2) - parameters -- etc. over the three age levels.

The picture found in item No. 5 shows how response categories are 'twisted' over the age levels, pointing to the fact, that 'agree' and 'not agree' are almost reverted concepts when collecting answers to this specific question - regarding different age groups. In order to understand this, one should keep in mind that analytic results obtained by means of the Rasch model **are results that assume single item information to be represented in the score** (four - or one-dimensional).

If you get other impressions from the simple distribution of **answers** given to the four response categories, it is exactly because such impressions do **not** at the same time take into consideration responses

given to other items of the dimensions. One can say that traditional statistical analyses like factor analysis are not 'modelled' to deal with such simultaneous information as explicit means in the process of analysis. On the other hand, all practical work based on questionnaire data is carried out by considering one, very simple way of combining simultaneous information kept in single-items, namely the **score total**.

Rasch models are mathematical 'solutions' to the description single-item behaviour **if** we require total score values to produce 'reasonable' results (i.e. still keeping full information concerning the pupil). One is, therefore, not all that free to evaluate test results of fit by **this model** if one really intends to go on by calculating total scores.

The described 'misfits' are consequently 'misfits' to the idea of taking total scores as instrumental variables in the further work. Not to say, that it is 'impossible' to calculate a total score - everybody can do that on a pocket calculator - but the calculated number is then just a number. It contains maybe little or 'no' information at all!

When we evaluate the appropriateness of the general Rasch model we are evaluating hypotheses that are **more** general than practical work assumes. Four-dimensional total scores (see Interpretions of tests) which are the reductions of single item information in this model are not suitable for practical work. One-dimensional total scores **are** the instrumental variables in practical work and strongest attention should, therefore, be given to results - e.g. sign of 'misfits' - obtained through the one-dimensional model.

If we follow these points of analytic strategy in case of DIM3 we proceed from the general model directly to results obtained by fitting (or rather 'not fitting') the one-dimensional model to data. The one-dimensional item powers of this model are for the two questions, item No. 3 and No. 5 analysed above:

	Item No. 3	Item No. 5
Age 8	0.22	-0.36
Age 11	-0.34	0.43
Age 14	-1.53	1.36

These figures cannot be fairly judged without considering, too, the scoring systems estimated within each age group (see above). However, they exibit another facet of 'twisted items' strong enough to be commented on. Roughly speaking, the sum of powers for these two items remains constant (near zero) for each age group. In terms of contributions to total scores it means, that whatever is 'lost' in agreement on the one item is 'gained' at another. This is like the situation in a series of simple mathematical tests (ability test), where two selected subtests may react opposite to two groups of pupils - the one test being difficult for one group, the other test being easy, and vice versa for the other group. Judged from total scores only, the two groups seem to be **equal**!

The one-dimensional Rasch model is rejected because DIM3 has 'twisted items', defined in this way. This indicates a **specific** insensibility of total score values, if total scores are calculated over these two items.

19. Factor analysis and homogeneity.

We shall briefly mention one aspect of applying Rasch models which calls for a traditional way of thinking - if this is conceived as concepts arising from factor analyses. The concept of 'loadings' from factor analysis leads one's mind towards 'importance', putting 'weights' to certain features, giving 'priority' to something inherent in a set of items.

These concepts are not present or defined in Rasch model analyses. However, the way total scores are built up in Rasch models can be compared to some extent with traditional analyses if we adopt a **conditional** way of thinking. Given the value of a total score we ask **how** it expectedly (i.e. by the mathematical expectation) must be composed under the model. Take $T_r = \Sigma x_{rj}$, the calculated score value as the sum of single item responses $x_{r1},\ldots,x_{rK}$ where we have K items in the dimension and x_{rj} being one of the scorings $\varphi(1)$, $\varphi(2)$, $\varphi(3)$ or $\varphi(4)$. We get the following relation equating total score T_r with **conditional expectations** of single responses $E(x_{rj} \mid T_r)$:

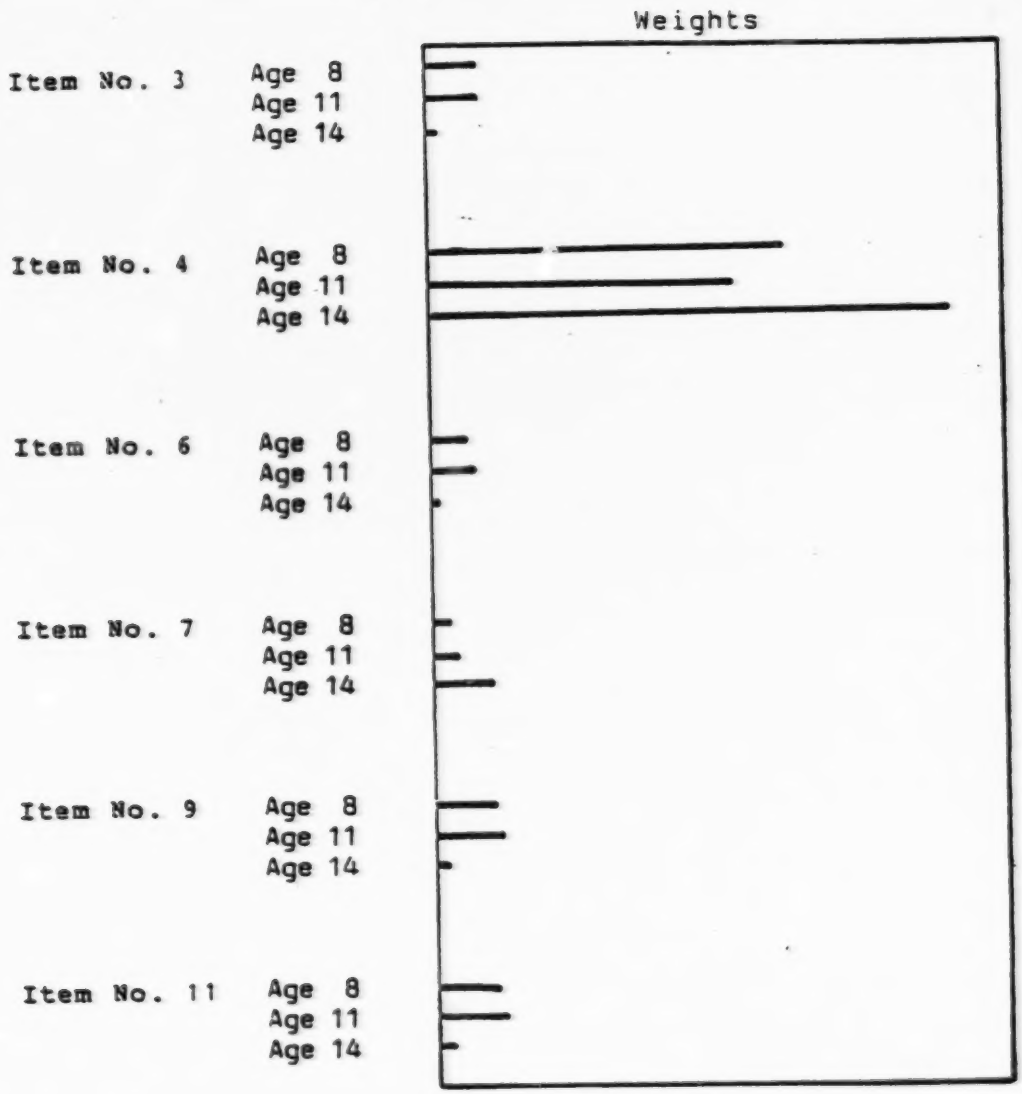
$$T_r = \sum_{(j)} E(x_{rj} \mid T_r)$$

$$= \sum_{j} \sum_{\mu} \varphi(\mu) \, \mathrm{Prob}(X_{rj} = \varphi(\mu) \mid T_r)$$

Total score T_r is therefore, a sum of (conditionally) expected answers $E(x_{rj} \mid T_r)$ $j = 1,\ldots,K$ given to the items. When 'misfits' occur by e.g. contrasting test results from girls and boys, one way of analysing the 'misfit' is to analyse how the two groups 'build up' score totals. Referring to the situation with mathematical tasks, we some think of pupils getting 'free' score points from very easy tasks - **easy for this group** - while other pupils considering the same tasks as generally more difficult 'lose' score points here. Equal score totals are in this case composed differently.

In DIM1 the situation as concerns the search for homogeneous items is rather confusing. We shall turn to this point later on, but first use DIM1 - items to illustrate the score-composing idea outlined above.

During one part of the analyses items Nos 3,4,6,7,9, and 11 are under evaluation for homogeneity, but numerical test results still show, that the model does not work properly, especially over age levels. We select two total score values: 1.0 which is a rather low value and 4.0 which is a rather high value. The range of possible values is from 0.0 to 6.0. 0.0 obtained through six zeros ($\varphi(4) = 0$) or six 'Not at all' responses and 6.0 obtained through complete 'A lot' agreements to all six items ($\varphi(1) = 1$).



Score=1.0

The items of ~~this subset~~ of DIM1 are

Item 3 (question 4): 'it is an easy thing for me to do'.

Item 4 (question 10): 'I get enjoyment from reading'.

Item 6 (question 19): 'it is fun to think I am a person in a story'.

Item 7 (question 25): 'it makes me happy'.

Item 9 (question 29): 'A book is my friend when I am lonely'.

Item 11 (question 41): 'It is nice to think about the things I have read'.

BEST COPY AVAILABLE

In the figure theweights or 'loadings' on item No. 3 - item No. 11 are pictured as columns, so that for each age group they sum up to 100%.

In table form we started from the expected values $E(X_{vi} \mid T_v = 1.0)$ within score group $T_v = 1.0$ (computer program for analysis of one-dimensional Rasch models, Allerup, 1985), and transform these values to fractions (percentage of T_v, cf the equation above).

DIM1 Item No.

		3	4	6	7	9	11	SUM
Score $T_v = 1.0$	Age 8	9	63	7	4	9	8	100%
	Age 11	9	53	8	5	13	12	100%
	Age 14	2	90	1	1	3	3	100%

Before interpreting the weight distribution displayed in the figure and table we calculate the weight distribution for the other total score group $T_v = 4.0$. Weight distribution for this score group is given below.

DIM1 Item No.

		3	4	6	7	9	11	SUM
Score $T_v = 4.0$	Age 8	17	23	14	10	17	17	100%
	Age 11	15	23	15	10	18	17	100%
	Age 14	16	21	13	12	17	18	100%

We know from the numerical tests of fit that the subset of items does **not** work as one homogeneous dimension. From tables and figures shown we now know **why**, and we can give explanations which are more detailled than the explanations given above, where deviations were just 'summed up' in item power estimates θ_i.

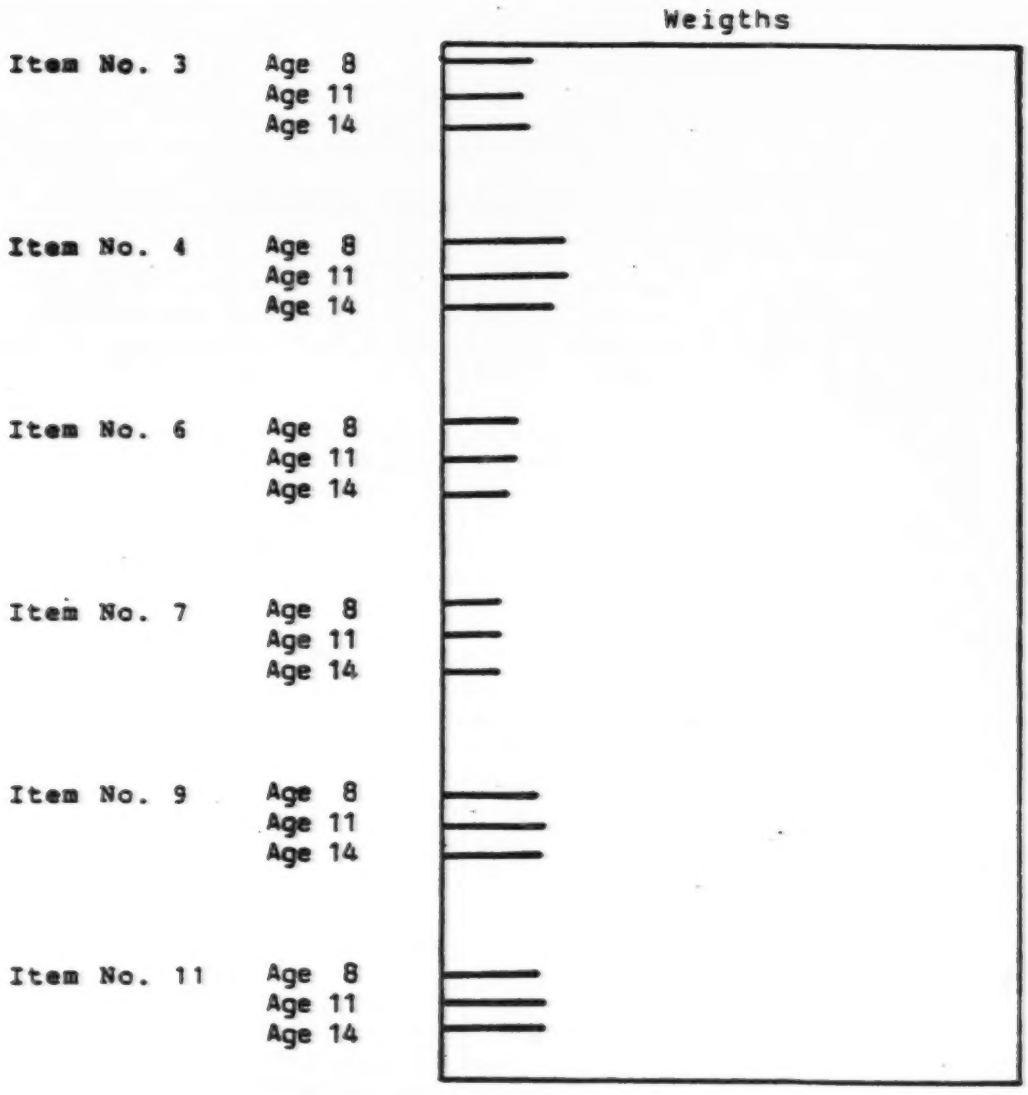
Greatest weights - i.e. attaching greatest 'importance' - are found with item No. 4: 'I get enjoyment from reading' whatever we deal with generally low-agreeing pupils ($T_v = 1.0$) or highly agreeing pupils ($T_v = 4.0$). This item 'delivers' the majority of 'positive agreement', any way counted in a total score value.

We conclude that 'factors' og weights are something inherent in Rasch models, too, but unlike genuine factor analyses in which factors are 'general' quantities connected to items, Rasch model 'factors' are quantities which are tied up with the basic conditional inference, i.e. their definitions are conditional upon total score T_v.

From an interpretational point oif view this is advantageous. In fact, it is seen from the analysis, that if you are generally very positive to the questions (score value 4.0, see below) the fraction of item 4 - contribution to total score is relatively small. If generally you do not agree with the questions (score 1.0) the only agreement left is hidden in item No. 4 - and here, for these quite negative pupils, we see substantial differences over age levels. Rather negative fourteen-year-old pupils give priority to 'enjoyment' and reject any thought of being 'a person in the story' (item No. 6).

BEST COPY AVAILABLE

DIM1



Score=4.0

20. Controlled and uncontrolled analyses.

Dim1 had in its primary stage of analysis twelve items. All items were selected at first hand by psychological and educational arguments only. The typical run of analysis is that first step analyses point to some few items being responsible for lack of fit of the model and after evaluation of the content, the formulation used the context seen in relation to other questions —etc. we may omit the question and proceed to second-step analyses and further on with a reduced subset of items.

It is not advisable to replace this sometimes cumbersome procedure by automatic stepwise procedures on a computer - not speaking of the possibility of letting a computer run **all** kinds of groupings of items! One special kind of 'automatization' may, however, be adequate. In cases of a suspicion, that **two** sub scales are built into the original set of items one may trigger the finding of such subdimensions by the following procedure.

With reference to the series of mathematical tasks once more we imagine one set of tasks suitable for girls, another suitable (i.e. Rasch model fits) for boys, but all tasks are entered into **one** ability test. 'Misfits' of the Rasch model for the **whole** set of tasks may then come out as indications of lack of homogeneity in the way, that some tasks are 'too easy' for girls (and difficult for boys), some other tasks being 'too easy' for boys (and difficult for girls).

On this pooled ability test the reverse problem of finding the two original subtests again is now solved by grouping the items (tasks) according to **direction** of deviation from homogeneity. In fact, tasks which are systematically too easy for girls go into one group while the remaining items go into the other group.

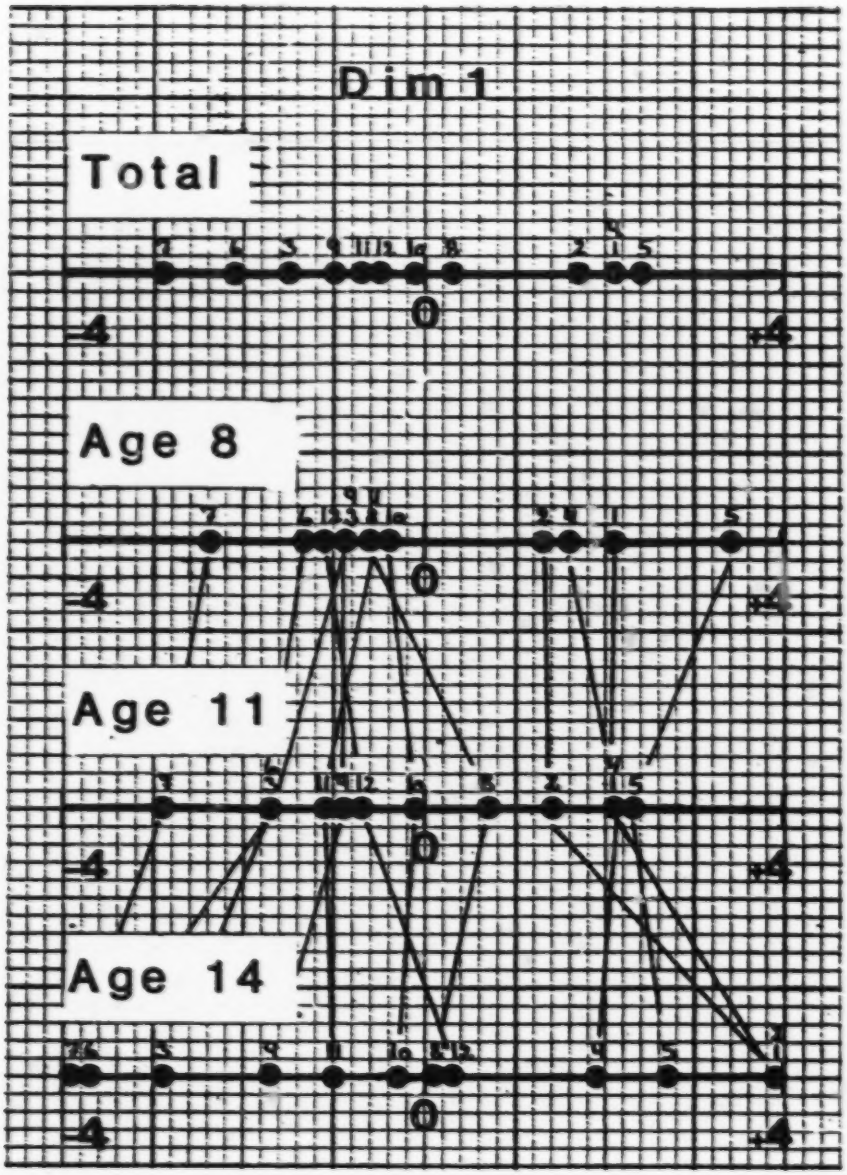
While this is a correct statistical solution to the problem of separating of one set of items into two homogeneous sub dimensions we now face another problem: How can mathematical ability be measured and compared with individuals of the other sex, when we use different rulers? But this is to be answered on another occasion.

In DIM1 the search for homogeneous items resembles the situation outlined above. It was not possible through several steps of analysis to draw such consequences of the test results that apart from 'few' items we had a homogeneous, reduced dimension. It appeared, however, from the sequence of tests that six items almost constantly turned 'left' over increasing age levels while the other six items turned 'right'. This systematism is pictured in the figure below, where the first step of analysis comprising all 12 items is shown.

Although not fully satisfied with numerical test results for test of the Rasch models we have indications so promising, that the original 12 items of DIM1 break down into two subdimensions each of which is composed of homogeneous items.

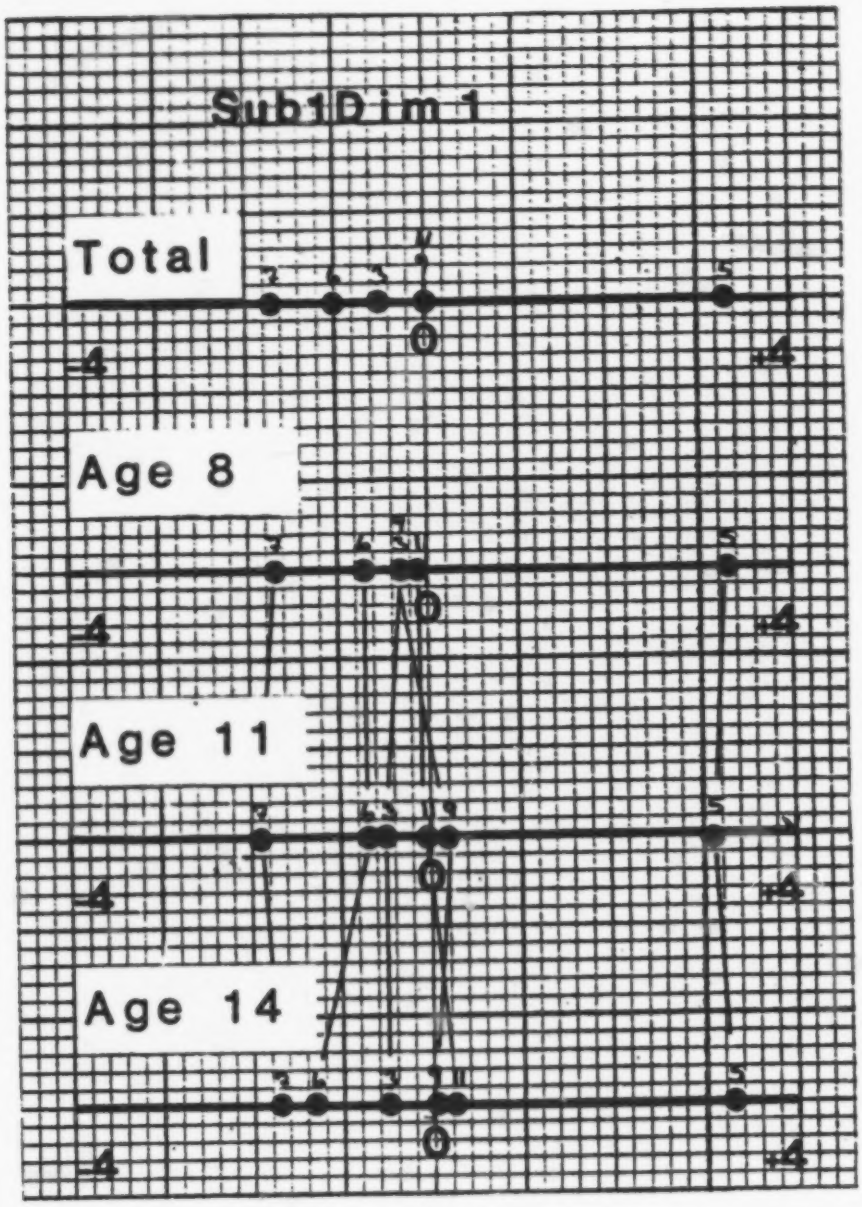
SUB1DIM1 : item Nos. 3,5,6,7,9 and 11

SUB2DIM1 : item Nos. 1,2,4,8,10 and 12.



The graphical representation of item power over age levels for SUBIDIM1 is given below, together with χ^2-test results for both general Rasch model and for the one-dimensional version

BEST COPY AVAILABLE



χ^2 (sex grouping | general model) = 23.9 df = 10

χ^2 (sex grouping | one-dimensional) = 16.8 df = 9

χ^2 (age grouping | general model) = 54.5 df = 20

χ^2 (age grouping | one-dimensional) = 44.8 df = 18

Work will be continued on the analysis of a.o. the two subdimensions. Consequently, we shall make no conclusive statements yet, but stress the background of their definitions: SUB1DIM1 consists of items which, **in company** with items from SUB2DIM1, have decreasing item powers (and vice versa for items of SUB2DIM1) when we study the change over increasing age levels.

BEST COPY AVAILABLE

A prudent interpretation of this fact goes along the line: As we increase age level

- excitement
- adventure
- enjoyment
- not forced to read
- when I am alone
- relaxation

which are SUB2DIM1 items, are receiving extra 'agreement' or attention at the cost of

- easy to do
- good story
- I am a person in the story
- makes me happy
- a book is my friend
- nice to think of the things read

which are SUB1DIM1 items.

Psychologically this is really not a strange result. We put emphasis on the conclusion, however, for different reasons, too. It demonstrates, how Rasch model analyses can be conducted so as to compete with factor analyses in **defining** 'factors', 'components' or dimensions! but with the basic difference that factor definitions take purely mathematical starting points in two-by-two item response correlations with **no** reference to total score values. Rasch model 'definition' of a dimension, refers in the first place explicitly to total score values as **the** means and, secondly, comes out of the analysis as results of testing **psychologically** rather than mathematically based hypotheses. In this respect factor analyses are likely to form rather than to test hypotheses concerning dimensions or factors.

21. Conclusions.

So far the analyses have been aimed at a critical view upon the fifty questions present in the questionnaire, in order to find consistent 'entries' to objective measurements. Although not all investigations came out successfully the future work will be based upon the working hypothesis that the original fifty questions can be grouped into the following seven dimensions, each of which meets the requirement that differentiated single-item information is collected by total score values calculated in accordance with proper scorings of the four response categories:

SUB1DIM1 :	Enjoyment-stimulation Nos. 4,13,19,25,29,41	6 items
SUB2DIM1 :	Enjoyment-fiction Nos.1,2,10,38,44	
DIM2 :	Furure Nos. 14,26,31,34,37,40,42,48	8 items
DIM3 :	Diversion Nos. 8,11,(21),45,49	(5) 4 items
DIM4 :	My country Nos. 39,47,50	3 items
DIM5 :	Professional skill Nos. 9,12,17,20,24,30	6 items
DIM6 :	Improvement of reading Nos. 22,33,36	3 items

Future steps of analysis will be statistical **and** educational analyses of the information collected by these seven dimensions - like collecting information by seven rulers. Each response pattern, originally compounded by a fifty - dimensional array of single item responses, will be compressed into a seven-dimensional array of measurements. We shall study how simultaneous 'scorings' on these seven dimensions may cast light upon 'why I like to read' and map age - and sex dependent characteristics with reference to these concepts.

22. References

Allerup P, Sorber G (1975): The Rasch model for questionnaires. Report No. 16. Danish Institute for Educational Research, Copenhagen.

Allerup P (1984): Analysis of developmental curves by means of Rasch Models. Proc. Symp. on applied statistics, Aarhus.

Allerup P (1985): One-dimensional Rasch models for questionnaires with more than two response categories (to appear from Educational Institute).

Andersen E. B.(1983): A general latent trait structure model. In: Wainer, H. and Messnick. S (ed) Principals of Modern Psychological Measurement. Hillsdale.

Andersen E. B. and Madsen, M (1976): Estimating the parameters of the latent population distribution. Report No. 35 Statistical Institute, University of Copenhagen.

Bech P, Allerup P, Gram L.F., Reisby N, Rosenberg R, Jacobsen O, Nagy A (1981): The Hamilton Depression Scale. Evaluation of objectivity using logistic models. Acta Psychiatry Scand. 63: 290-299.

Bech P., Allerup P, Rosenberg R (1978): The Marke Nymann Temperament Scale. Acta Psychiatr. Scand. 57: 49-58.

Bech P, Allerup P (1984): Assessment of symptom change from improvement curves on Hamilton depression Scale in trials with antidepressants, Psychopharmacology 84: 276-281.

Fischer G (1974): Einführung in Theorie psychologischer Tests. H. Huber, Vienna.

Greaney V. and S. Neumann (1984): Exploring the functions of reading: An International Perspective. Paper read at 10th congress on Reading. Hong Kong, 1984.

Rasch G. (1961): On general laws and the meaning of measurement in psychology, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Tome 4, pp. 321-333.

Rasch G. (1966): An individualistic approach to item analysis. In: Lazarsfeld, P.F. & N.W. Henry (eds) Reading in Mathematical Social Science, Science Research Associates, Chicago.

Rasch G. (1966): An informal report on the present state of a theory of objectivity in comparisons, Proceedings of the NUFFIC International Summer Session in Science at "Het Onde Hof", The Hague.

Rasch G (1968): A mathematical theory of objectivity and its consequences for model construction. European Meeting on Statistics, Econometries and Management Science, Amsterdam.

Rasch G (1977): On specific objectivity. Danish Yearbook of Philosophy, vol. 14, Munksgaard, Copenhagen, pp 58-94.

Rasch G (1980): Probalistic models for some intelligence and attainment tests. Danish Institute for Educational Research, Copenhagen 1960. (Reprinted by University of Chicago Press, Chicago 1980).

23. Appendix 1.

- The questionnaire.

Question No.

1. I read because it is exciting
2. I read because it makes me feel like I'm going into another world and having an adventure
3. I read because it helps me at school
4. I read because it is an easy thing for me to do
5. I read because it is interesting
6. I read to know more about my religion
7. I read because it helps me with subjects I will study later on
8. I read because it gets my mind off things
9. I read because it makes me understand more
10. I read because I get enjoyment from reading
11. I read because it stops me from being bored
12. I read because it puts thoughts in my mind
13. I read because books tell good stories
14. I read because it will help me get a good job
15. I read because my parents think it is important
16. I read because it helps me forget about my worries
17. I read so I can learn how people live in other countries

18. I read so I can go to college

19. I read because it is fun to think I am a person in a story

20. I read so I will learn how to respect other people

21. I read because it helps me go to sleep easily

22. I read because it will help me pass my examinations

23. I read when I am lonely

24. I read because it will help me learn something new

25. I read because it makes me happy

26. I read because it helps me to think of a better world

27. I read when there is nothing exciting on TV

28. I read when I'm not forced to read

29. I read because a book is my friend when I am lonely

30. I read so I can imagine what it's like at another time in history

31. I read because it will help me to work better

32. I read because it will help me know what is right and what is wrong

33. I read because people will praise me for being a good reader

34. I read because books show me the right way to live

35. I read because it will help me to become an important person

36. I read because it will help me to learn more words

37. I read because it will help me to live better

38. I read because it is a nice thing to do when I am alone

39. I read because it will help my country if I learn things

40. I read so that when I grow up I will know the world better

41. I read because it is nice to think about the things I have read

42. I read because it will help me to improve what I write

43. I read because it is something I can do as quickly or as slowly as I like

44. I read because it helps me to relax

45. I read because it makes the time go by

46. I read because my teacher thinks it is important

47. I read because the government in my country wants me to be able to read

48. I read so that I can work better

49. I read because I have nothing else to do

50. I read so I will know how to help my country.

The Danish Institute
for Educational
Research

28 Hermodsgade
DK 2200
Copenhagen N
Denmark

Phone
Internat.:
+ 45 1 81 01 40

The Danish Institute for Educational Research is engaged in educational research, educational development and informative work in education. Studies can be related to all age-groups - from pre-school to retirement.

The Danish Institute for Educational Research was established in 1955 and is directly responsible to the Ministry of Education. It is closely concerned with all aspects of educational activity and co-operates with other scientific institutes and individuals to an extent determined by the projects and studies.

The objects of the Institute are as follows:

a. to carry out scientific research and studies of importance to educational activities.
b. to assist in developmental projects for which scientific support is required and to offer advisory service, planning, collection and analysis of educational material.
c. to provide students of psychology and education with trainee facilities at the Institute.
d. to publish reports of its activities and provide information in general, relating to matters of interest in the field of education.
e. to construct and distribute tests and similar material which by their nature cannot be made freely available.

The management of the Institute consists of **the Director** who conducts the daily administrative and scientific work, and **the Research Board** composed of the Director, **the Research Directors** of the Departments of the Institute, and the following members appointed by the Minister of Education: 8 university professors or other university lecturers, 1 school hygienist, and one member of the Institute's Council of Representatives, nominated by the Council.

The Research Board is the highest authority of the Institute in all scientific matters. It is the duty of the Research Board to approve the Institute's research programme, to follow developments in educational and psychological research, and to initiate research projects. The Research Board holds 10 annual meetings.

The Institute has a **Council of Representatives**, composed of representatives of educational organisations, other professional associations and institutions within the Institute's field of activities. The Council of Representatives serves as a contact between the Institute and the various fields of education. The Council is entitled to discuss all matters concerning the Institute and to make recommendations to the management of the Institute. It usually meets once a year.

At the Institute the following **departments, etc.** have been established:
Dept. 1. Department for research in educational planning and action
Section for differential pedagogics
Dept. 2. Department for basic research in educational psychology
Dept. 3. Department for tests
Dept. 4. Department of educational experiments
Dept. 5. Department for statistics (with computer terminal)
Laboratory (ITV etc.)
Administration office (with sales office).

Publication of work on research and development, etc.
A list of the publications can be found in the **annual lists of publications** (in Danish).

Moreover, an annual **report** is published which also contains the research programme of the Institute.

A description of the Institute can be found in »Danmarks pædagogiske Institut. Beretning 1977-78. Anden del. En undersøgelse af Danmarks pædagogiske Instituts funktion«. (An analysis of the function of the Institute).

Current information on i.a. projects and new publications is given in »NYT fra Danmarks pædagogiske Institut« which is published regularly in Danish.

Information and advisory service are offered through the staff's membership of boards, councils, commissions and committees, work groups, etc. in the field of education, and in addition the Institute assists in the working up of material, answers enquiries and assists with advice in the field of educational development.

The Institute was established under the Act of 11th June, 1954, and the rules governing the Institute have been laid down in the Executive Order of 2nd March, 1965.

Vend venligst.

BEST COPY AVAILABLE

Danmarks
Pædagogiske
Institut

Hermodsgade 28
2200 København N
Tlf. (01) 81 01 40

Danmarks pædagogiske Institut arbejder med forsknings- og udviklingsarbejde inden for opdragelse og uddannelse. Arbejdet kan foregå inden for alle livsaldre fra førskolealderen til pensionistalderen.

Danmarks pædagogiske Institut er oprettet i 1955. Det hører direkte under undervisningsministeriet med nøje samarbejde til alle sider af pædagogisk virksomhed. Det samarbejder med andre videnskabelige institutter og enkeltpersoner i det omfang, de enkelte arbejder og undersøgelser forudsætter.

Instituttet har følgende **formål:**

a. At foretage undersøgelser og udføre arbejder af betydning for pædagogisk virksomhed.
b. At assistere i pædagogisk forsøgsvirksomhed, hvortil videnskabelig støtte ønskes, og at være behjælpelig med rådgivning, planlægning og indsamling og bearbejdelse af materialer.
c. At give psykologi- og pædagogikstuderende adgang til praktiktjeneste ved instituttet.
d. Gennem publikationer at gøre rede for resultaterne af sit arbejde og i øvrigt at sørge for orientering om forhold af væsentlig betydning for pædagogisk virksomhed.
e. At udarbejde og forestå udsendelse af testmateriale o.lign., som på grund af sin beskaffenhed ikke kan være frit tilgængeligt.

Instituttets **ledelse** består af direktøren, der har den daglige administrative og faglige ledelse, og af et fagligt råd bestående af direktøren, lederne af instituttets afdelinger og følgende af undervisningsministeren beskikkede medlemmer: 8 professorer eller andre universitetslærere, 1 skolehygiejniker og 1 medlem af instituttets repræsentantskab efter indstilling fra repræsentantskabet.

Det faglige Råd er instituttets højeste myndighed i alle videnskabelige spørgsmål. Det påhviler rådet at godkende instituttets arbejdsplan, at følge udviklingen inden for det pædagogiske og det pædagogisk-psykologiske forskningsområde og at tage initiativ til gennemførelse af forskningsopgaver. Rådet holder normalt 10 møder om året.

Til instituttet er knyttet et **repræsentantskab** sammensat af repræsentanter for pædagogiske organisationer, andre faglige foreninger og institutioner inden for instituttets arbejdsområder. Repræsentantskabet formidler kontakt mellem instituttet og de pædagogiske arbejdsområder. Det kan drøfte alle spørgsmål vedrørende instituttet og gøre indstilling herom til instituttets ledelse. Det holder normalt møde én gang årligt.

Ved instituttet er oprettet følgende **afdelinger m.v.:**
Afdeling 1. Afdelingen for didaktiske undersøgelser
Sektion for differentiel pædagogik
Afdeling 2. Afdelingen for pædagogisk-psykologisk grundforskning
Afdeling 3. Afdelingen for tests og prøver
Afdeling 4. Afdelingen for pædagogisk forsøgsvirksomhed
Afdeling 5. Afdelingen for statistik (herunder edb-terminal)
Laboratorium (ITV m.m.)
Administrationskontor (med salgsafdeling)

Formidling af forsknings- og udviklingsarbejde m.v.
Instituttet udgiver publikationer om foretagne forsknings- og udviklingsarbejder m.v.

En fortegnelse over dem findes i de **årlige publikationsoversigter.**

Desuden udgives der hvert år en **beretning,** som også indeholder instituttets arbejdsplan.

En beskrivelse af instituttet findes i »Danmarks pædagogiske Institut. Beretning 1977-78. Anden del. En undersøgelse af Danmarks pædagogiske Instituts funktion«

Løbende orientering om bl.a. arbejder og nye publikationer gives i »NYT fra Danmarks pædagogiske Institut«, der udsendes gratis.

Oplysende og rådgivende virksomhed udføres gennem institutmedarbejdernes medlemskab af råd, nævn, kommissioner, udvalg og arbejdsgrupper m.v. inden for opdragelses- og uddannelsesområderne. Hertil kommer assistance ved bearbejdelse af materialer, ved besvarelse af forespørgsler og ved rådgivende assistance til pædagogisk udviklingsarbejde.

Instituttet er oprettet i henhold til lov af 11. juni 1954, og reglerne for instituttet er fastsat ved bekendtgørelse af 2. marts 1965.

Please turn over

BEST COPY AVAILABLE

Un 01 18-193-6